MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABDELHAMID BEN BADIS DE MOSTAGANEM
FACULTÉ DES SCIENCES EXACTES ET DE L'INFORMATIQUE

THÈSE

Doctorat LMD

pour obtenir le grade de Docteur délivré par

Université de Mostaganem

Spécialité "Optimisation, Ondelettes et Calcul Fractionnaire"

présentée et soutenue publiquement par

Mme. Naima DJELLOUL Epouse LATROCH

le 03 juillet 2018

Optimisation Quadratique pour les Machines à Vecteurs de Support (SVMs)

Directeur de thèse : **Abdessamad AMIR (Prof. Université Abdelhamid Ibn Badis, Mostaganem)**

Jury

Examinateur,	Djamel BENTERKI	Prof. Université Abbas Ferhat, Setif1.
Examinateur,	Omar BELHAMITI	Prof. Université Abdelhamid Ibn Badis, Mostaganem.
Président,	Zoubir DAHMANI	Prof. Université Abdelhamid Ibn Badis, Mostaganem.

LABORATOIRE DE MATHÉMATIQUES PURES ET APPLIQUÉES
FACULTÉ DES SCIENCES EXACTES ET DE L'INFORMATIQUE (FSEI)
Chemin des Crêtes (Ex-INES), 27000 Mostaganem, Algérie

THESE

Remerciements

Je remercie Dieu pour m'avoir armer de force, de patience, de volonté et de courage durant mes années de thèse.

Mes plus sincères remerciements s'adressent à : le professeur Abdessamad AMIR. Au travers de nos discussions, il m'a apporté une compréhension plus approfondie des divers aspects du sujet. Je salue aussi sa souplesse, son ouverture d'ésprit et sa bonne humeur qui ont su me laisser une large marge de liberté pour mener à bien ce travail de recherche.

Je désire également exprimer ma profonde gratitude au professeur Zoubir DAHMANI qui m'a honoré en acceptant d'être président de ce jury ainsi qu'aux professeurs Djamel BENTERKI et Omar BELHAMITI qui m'ont honoré d'être examinateurs de ce jury.

J'ai une pensée toute particulière à ma très proche et chère enseignante Mme. ABLAOUI.
Ma gratitude va de plus à mes amies Kheira, Louiza,Nadira, Leila et Nawel. Enfin, je ne saurais terminer cette liste sans adresser un remerciement particulier à Zineb KISSERLI.

Je ne voudrais pas oublier mes beaux-parents et mes collégues de travail.

Ma plus grande gratitude et reconnaissance vont bien sûr à mes parents pour m'avoir offert la possibilité d'effectuer mes études dans les meilleurs conditions. Pour leur soutient permanent, que mes frères : Fayçal, Nacer et Hamza, ma soeur : Hadjer et toute ma famille, trouvent ici toute ma reconnaissance.

Je remercie enfin et surtout ma petite famille : mon mari Ali et mon trésor Mahieddine.

Table des matières

Table des figures

Notations

- $\mathbb{N}$: l'ensemble des entiers naturels.
- $\mathbb{R}$: l'ensemble des nombres réels.
- $\mathbb{C}$: l'ensemble des nombres complexes.
- x_i : la $i\_ème$ composante du vecteur x.
- x^j : le $j\_ème$ vecteur de la matrice A.
- $f \in \mathscr{C}^1$: f est continuement dérivable.
- rg(A) : le rang de la matrice A.
- $\nabla f(x)$: le gradient de la fonction f.
- $\langle .,. \rangle$: Produit scalaire.
- ℓ_2 : l'espace des suites de carrés sommable.
- L_2 : l'espace des fonctions de carré intégrable.
- δ_{ij} : symbole de Kronecker.
- $\mathbb{R}^{\Omega}$: l'ensembles de toutes les fonctions définies sur $\Omega \subset \mathbb{R}^d$ à valeurs réelles.
- $\mathrm{H}ilb(\mathbb{R}^{\Omega})$: l'ensemble de tous les RKHS de $\mathbb{R}^{\Omega}$.
- $\mathrm{A} \subset \mathbb{R}^{m \times n}$: A matrice à coéfficients réels, à m lignes et n colonnes.
- ker A : noyau de la matrice A.
- Im A : image de la matrice A.
- $\overline{z}$: le conjugué du nombre complexe z.
- $span\{x_1, \cdots, x_n\}$: l'espace vectoriel engendré par les vecteurs $x_1, \ldots, x_n$.
- $\overset{\circ}{\mathrm{X}}$: l'intérieur de l'ensemble X.

Introduction

Les machines à vecteurs de support (SVM) sont un ensemble d'algorithmes d'apprentissage supervisé utilisées pour la classification et la régression. Les algorithmes SVMs ont été largement appliqués en bioinformatique, biologie, traitement du signal EEG, reconnaissance de formes et autres problèmes du monde réel (Wang2005 [47] et Spiliopoulou et al. [38]). Nous nous concentrons ici sur un type particulier de problèmes d'apprentissage, à savoir la classification binaire, lorsque les données ont deux classes. Les SVMs classent les données en trouvant la meilleure fonction de décision, qui sépare tous les points de données d'une classe de ceux de l'autre classe [10]. Les SVM sont basés sur trois concepts mathématiques :

- La théorie de la généralisation : la formulation utilise le principe de minimisation des risques structurels (SRM) (Vapnik [44]), qui s'est révélé supérieur au principe traditionnel de minimisation des risques empiriques (ERM), utilisé par les réseaux neurones conventionnels. (Burge1998 [5]).

- La théorie de l'optimisation : Le coprs des SVMs est un programme quadratique convexe (QP), dont l'analyse mathématique et la résolution numérique ont connu un développement considérable ces dernières années ([30]), des solveurs efficaces sont maintenant disponibles afin de résoudre ces problèmes même avec des données de grande taille et denses (Survit et all.2012[41]).

- Astuce noyau : L'ensemble d'apprentissage n'est généralement pas linéairement séparable dans l'espace d'entrée, les fonctions noyaux sont utilisées pour transformer les données d'entrée dans un espace à grande dimension nommé espace caractéristique. Basé sur les espaces de Hilbert à noyau reproduisant (RKHS) et le théorème de Mercer (Aronszajn1950 [1]), le noyau défini un produit scalaire dans l'espace caractéristique pour lequel l'image de l'ensemble d'apprentissage soit linéairement séparable.

Plusieurs fonctions noyaux sont maintenant disponibles, telles que les noyaux linéaires, polynomiales, Sigmoid et le noyau de base radial (RBF). Grâce à ses propriétés théoriques et numériques, le noyau RBF est le plus utilisé en pratique. En fait, RBF est un noyau universel (Micchelli et all.2006[28]) et efficacement calculable. Les noyaux universels ont la propriété géométrique surprenante de séparer l'image de tous ensembles disjoints A, B dans un espace métrique compact X. Ceci est une conséquence du fait que son RKHS est dense dans l'ensemble des fonctions continues sur X noté par C(X). Concevoir une fonction de noyau universelle avec un bon comportement numérique est en effet un grand défi dans le domaine de l'apprentissage automatique.
Les noyaux pour les SVM peuvent être construits de plusieurs manières. Récemment, une attention particulière a été accordée à la construction basée sur les concepts de poly-

nômes orthogonaux qui ont été largement utilisés dans tous les domaines de la science et de l'ingénierie. Zhou et al.2007 ([50]) ont proposé une nouvelle approche pour construire une fonction générale de noyau appelée noyau polynomial de Jacobi. Ye et al ([49]) ont accordé plus d'attention au noyau polynomial de Chebyshev et Ozer et al.2011 ([31]) l'ont étendu au noyau de Chebyshev généralisé. V. H. Moghaddam et J. Hamidzadeh ([29]) ont étudié le noyau polynomial d'Hermite et ils ont proposé de nouvelles fonctions de noyau qui sont construites en combinant les fonctions de noyau d'Hermite et de noyaux universels. Le noyau de Legendre a également été considéré par Pan et al.2012 ([32]) et comparé au noyau de Chebyshev dans le domaine de segmentation d'image. Notre but dans ce travail est de caractériser les propriétés théoriques possédées par le noyau de Legendre. En utilisant la théorie des produits tensoriels (Ryan2002[36]), nous identifions le noyau RKHS de Legendre sur un ensemble de données multidimensionnel, ce qui nous permet de comprendre sa capacité à extraire les caractéristiques les plus discriminantes et de montrer que ce noyau a de puissantes propriétés de séparation. Ces résultats sont confirmés et supportés par des expériences numériques.

Le manuscrit s'articule au tour de quatre chapitres. Le premier chapitre traite les problèmes d'optimisation quadratique, caractérisation de la solution optimale et la dualité et aussi quelques algorithmes de résolution ad-hok. Le second chapitre introduit le prblème des SVMs, commençant par le cas linéairement séparable, au cas non linéairement séparable. Quant au troisième chapitre présentra une introduction aux fonctions noyaux, utiles pour l'étude des SVMs. Le quatrième chapitre résume notre contribution en caractérisant les proprétés théoriques possédées par les polynômes orthogonaux de Legendre. En utilisant la théorie du produit tensoriel, nous identifions le noyau RKHS de Legendresur un ensemble de données multidimensionnel, ce qui nous permet de comprendre sa capacité à extraire les caractéristiques les plus discriminatives. Ces résultats sont confirmés et supportés par des expériences numériques.

Chapitre 1

Programmation quadratique

Un programme non linéaire avec contraintes, s'écrit d'une manière générale sous la forme suivante :

$$\begin{cases} \min f(x) \\ h_j(x) = 0\,,\ j = 1, \cdots, m \\ g_i(x) \leqslant 0\,,\ i = 1, \cdots, p \\ x \in \mathrm{X} \subset \mathbb{R}^n, \end{cases} \tag{1.1}$$

où $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ est une fonction différentiable, non nécessairement linéaire appelée fonction objectif, $\mathrm{S} = \left\{x \in \mathrm{X} \subset \mathbb{R}^n,\ h_j(x) = 0,\ j = 1, \cdots, m \text{ et } g_i(x) \leqslant 0,\ i = 1, \cdots, p\right\}$ est un sous ensemble de $\mathbb{R}^n$, appelé ensemble des solutions admissibles ou réalisables.

Exemple 1.1

$$\begin{cases} \min f(x) = (x_1 - 1)^2 + x_2 - 2 \\ x_2 - x_1 = 1 \\ x_1 + x_2 \leqslant 2. \end{cases} \tag{1.2}$$

Le probème est sous sa forme standard avec $f(x_1, x_2) = (x_1 - 1)^2 + x_2 - 2$ *et* $\mathrm{S} = \left\{x \in \mathbb{R}^2 / g(x) = x_1 + x_2 \leqslant 2, h(x) = x_2 - x_1 = 1\right\}$ *ce probème peut être résolu graphiquement. Dans la figure (1.1), la région admissible est la demi droite tracée en gras et les paraboles représentent les niveaux de la fonction objectif* f*. On remarque que les ensembles de niveaux de* f*,rencontrent la demi droite en premier lieu au point* x^**, au delà, la fonction prend des valeurs supérieurs. D'où la solution optimale du problème (1.2) est* $x^* = \left(\frac{1}{2}; \frac{3}{2}\right)^\top$ *et la valeur optimale associée est* $f(x^*) = \frac{-1}{4}$*. Il est claire qu'au delà de 4 variables (*$n > 3$*), le problème (1.1) ne peut être résolut géométriquement, d'où la nécessité de développer des méthodes numériques de résolution, qui s'appuient sur une analyse mathématique rigoureuse.*

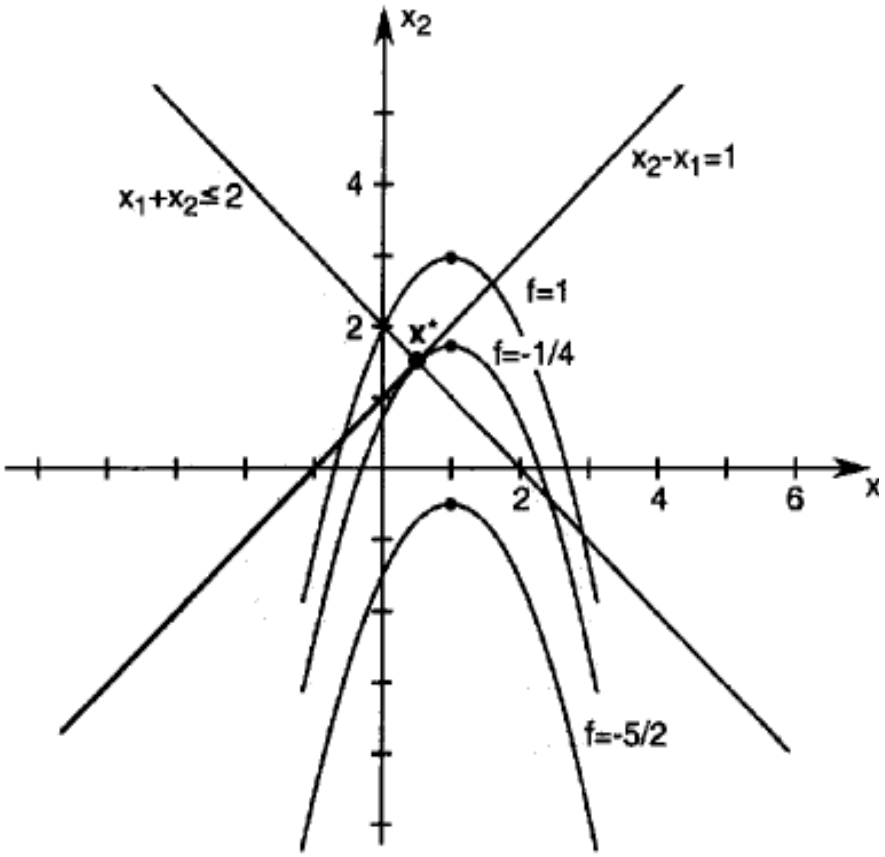


FIGURE 1.1 – Résolution graphique de l'exemple 1.1[12]

1 Contraintes égalités

La résolution du problème (1.1) passe souvent par la résolution d'un programme non linéaire avec contraintes égalités, qui s'écrit sous la forme suivante :

$$\begin{cases} \min f(x) \\ h_j(x) = 0, j = 1, \cdots, m \\ x \in \mathrm{X} \subset \mathbb{R}^n \end{cases} \tag{1.3}$$

Les contraintes peuvent être exprimées sous la forme $h(x) = 0$ avec $h(x) = [h_1(x), h_2(x), \cdots, h_m(x)]^\top$ avec $m \leqslant n$. La fonction h est supposée au moins de classe $\mathscr{C}^1$.

Définition 1.1 *[3] Un point $\overline{x}$ satisfaisant les contraintes $h_1(\overline{x}) = 0$, $h_2(\overline{x}) = 0, \cdots, h_m(\overline{x}) = 0$ est dit un point régulier des contraintes, si la famille de vecteurs gradient*

$$\{\nabla h_j(\overline{x})\}_{j=1,\cdots,m}$$

sont linéairement indépendants, ou d'une manière équivalente, la matrice Jacobienne

$$\mathrm{D}h(\overline{x}) = \left(\nabla h_1(\overline{x})^\top, \cdots, \nabla h_m(\overline{x})^\top \right)$$

est de rang maximal, c-à-d : si $m \leqslant n$ alors $rg(\mathrm{D}h(\overline{x})) = m$.

Le théorème suivant donne une condition nécessaire d'ordre 1.

Théorème 1.1 (Théorème de Lagrange) *[12] Soit x^* un minimum local (ou maximum local) du problème (1.3). Supposons que x^* est un point régulier. Alors, il existe $\lambda^* \in \mathbb{R}^m$ tel que :*

$$\nabla f(x^*) + \mathrm{D}h(x^*)^\top(\lambda^*) = 0$$

ou d'une manière équivalente :

$$\nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla h_j(x^*) = 0.$$

Dans le cadre de la convexité, on peut avoir des conditions nécessaires et suffiantes uniquement d'ordre 1.

Définition 1.2 *[3] Soit la fonction $f : \mathrm{C} \subset \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$, avec C un ensemble convexe, f est dite convexe si pour tout $x, y \in \mathrm{C}$ et pour tout $\lambda \in [0,1]$*

$$f((1-\lambda)x + \lambda y) \leqslant (1-\lambda) f(x) + \lambda f(y).$$

Le théorème suivant donne deux caractérisations des fonctions convexes. La première via un ensemble convexe, la deuxième via le gradient.

Théorème 1.2 *[23] Soit la fonction $f : \mathrm{C} \subset \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$, avec C un ensemble convexe.*

1. *f est convexe si et seulement si l'épighraphe de f noté $epi f$ est convexe, avec*

$$epi f = \left\{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} : \ f(x) \leqslant \alpha\right\}. \tag{1.4}$$

2. *Si f est différentiable et C ouvert, alors f est convexe si et seulement si :*

$$\forall x, y \in \mathrm{C} \ : \ f(y) \geqslant f(x) + \nabla f(x)^\top (y - x).$$

Le théorème suivant donne la caractérisation désirée.

Théorème 1.3 *[12] Soit la fonction $f : \mathrm{X} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$, f est convexe et de classe $\mathscr{C}^1$ sur un ouvert qui contient l'ensemble convexe des solutions admissibles.*

$$\mathrm{S} = \left\{x \in \mathbb{R}^n : \ h(x) = 0\right\}$$

avec $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, $h \in \mathscr{C}^1$. Alors x^* est un minimum global de S si et seulement s'il existe $\lambda^* \in \mathbb{R}^m$ tel que

$$\nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla h_j(x^*) = 0.$$

2 Contraintes mixtes (égalités et inégalités)

Rappelons le problème général avec contraintes mixtes avec $\mathrm{X} \subset \mathbb{R}^n$

$$\begin{cases} \min f(x) \\ h(x) = 0, \\ g(x) \leqslant 0, \\ x \in \mathbb{R}^n, \end{cases} \tag{1.5}$$

avec $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ et $g : \mathbb{R}^n \longrightarrow \mathbb{R}^p$.

Définition 1.3 *[43] Une contrainte d'inégalité $g_i(x) \leqslant 0$, $i = 1, \cdots, p$ est dit **active** en x^* si $g_i(x^*) = 0$. Elle est inactive en x^* si $g_i(x^*) < 0$.*

Soit x^* satsfaisant $h(x^*) = 0$ et $g(x^*) \leqslant 0$. Soit l'ensemble des indices des contraintes actives $\mathrm{J}(x^*)$ avec

$$\mathrm{J}(x^*) = \left\{i, g_i(x^*) = 0\right\}.$$

Alors, on dit que le point x^* est régulier si les vecteurs

$$\nabla h_j(x^*),\ \nabla g_i(x^*),\ 1 \leqslant j \leqslant m,\ i \in \mathrm{J}(x^*)$$

sont linéairement indépendants.

Théorème 1.4 (Karuch Kuhn and Tucker (K-K-T)) *[43] Soient f,h et $g \in \mathscr{C}^1$. Soit x^* un point régulier et un minimum local du problème (1.5).*

Alors, il existe $\mu^* \in \mathbb{R}^m$ et $\lambda^* \in \mathbb{R}^p$ tel que :

1. $\lambda^* \geqslant 0$.
2. $\nabla f(x^*) + \sum_{j=1}^{m} \mu_j^* \nabla h_j(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla g_i(x^*) = 0.$
3. $(\lambda^*)^\top g(x^*) = 0.$

Dans ce cadre aussi la convexité donne des conditions nécessaires et suffisantes

Théorème 1.5 *[43] Considérons le problème d'optimisation (1.5), avec $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ et $g : \mathbb{R}^n \longrightarrow \mathbb{R}^p$ des fonctions de classe $\mathscr{C}^1$. Supposons que f et g sont convexes et h est affine. Supposons aussi qu'il existe x^* admissible et qu'il existe $\mu^* \in \mathbb{R}^m$ et $\lambda^* \in \mathbb{R}^p$ tel que*

1. $\lambda^* \geqslant 0$
2. $\nabla f(x^*) + \sum_{j=1}^{m} \mu_j^* \nabla h_j(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla g_i(x^*) = 0$
3. $(\lambda^*)^\top g(x^*) = 0.$

Alors, x^* est un minimum global du problème (1.5).

3 La dualité Lagrangienne

Reconsidérons le problème (1.1), que nous allons appeler ici problème primal. Plusieurs problèmes étroitement liés à ce problème ont été proposés dans la littérature et ils sont appelés problèmes duaux. Parmi les différentes formulations de dualité, la dualité Lagrangienne a peut être attiré le plus d'attention. Elle a conduit à plusieurs algorithmes pour résoudre des problèmes de grande taille. Le problème dual Lagrangien est établit ci-dessous.

Définissons le Lagrangien

$$\mathscr{L}(x, \lambda, \mu) = \left\{f(x) + \lambda^\top g(x) + \mu^\top h(x) : x \in \mathrm{X}\right\}, \tag{1.6}$$

et la fonction

$$\theta(\lambda, \mu) = \inf_x \mathscr{L}(x, \lambda, \mu), \tag{1.7}$$

appelé fonction dual et définissons le problème

$$\begin{cases} \max \theta(\lambda, \mu) \\ \lambda \geqslant 0. \end{cases} \tag{1.8}$$

Ce problème est appelé le problème dual associé au problème(1.1).

Théorème 1.6 (La dualité faible) *[3] Soit $\overline{x}$ une solution admissible du problème (1.1) et $(\overline{\lambda},\overline{\mu})$ est admissible pour le problème (1.8). Alors*

$$f(\overline{x}) \geqslant \theta(\overline{\lambda},\overline{\mu}).$$

Théorème 1.7 (La dualité forte) *[3] Soient* X *un ensemble non vide convexe de* $\mathbb{R}^n$, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $g : \mathbb{R}^n \longrightarrow \mathbb{R}^p$ *convexes et* $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *affine (i.e., de la forme* $h(x) = Ax - b$*). Supposons qu'il existe* $\widehat{x}$ *tel que* $\widehat{x} \in X$, $g(\widehat{x}) < 0$, $h(\widehat{x}) = 0$ *avec* $0 \in \overset{\circ}{h}(X)$ *et* $h(X) = \{h(x) \setminus x \in X\}$. *Alors*

$$\inf\{f(x) : x \in X,\ g(x) \leqslant 0,\ h(x) = 0\} = \sup\{\theta(\lambda,\mu),\ \lambda \geqslant 0\}.$$

De plus, si inf *est fini, alors le* sup *est atteint en un point* $(\overline{\mu},\overline{\lambda})$ *avec* $\overline{\lambda} \geqslant 0$. *Si* inf *est atteint au point* $\overline{x}$, *alors* $\overline{\lambda}^\top g(\overline{x}) = 0$.

Bien qu'on peut voir les propriétés de la dualité lagrangienne sur des exemples plus généraux ; dans le théorème (1.7) les fonctions f et g ne sont pas forcément différentiables, on va se restreindre dans cette étude seulement au cas quadratique.

On considère d'abord le problème quadratique suivant

$$\begin{cases} \min \frac{1}{2}x^\top Qx + c^\top x \\ Ax \leqslant b, \end{cases} \tag{1.9}$$

avec $Q \in R^{n\times n}$ une matrice symétrique et semi-définie positive, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ et $b \in \mathbb{R}^m$. La fonction dual est définie par

$$\theta(\lambda) = \inf\left\{\frac{1}{2}x^\top Qx + c^\top x + \lambda^\top(Ax - b) : x \in \mathbb{R}^n\right\}, \tag{1.10}$$

et le problème dual est

$$\begin{cases} \max\theta(\lambda) \\ \lambda \geqslant 0. \end{cases} \tag{1.11}$$

Notons que pour un λ donné la fonction $\frac{1}{2}x^\top Qx + c^\top x + \lambda^\top(Ax - b)$ est convexe, d'où une condition nécessaire et suffisante d'optimalité est donnée par

$$Qx + A^\top\lambda + c = 0. \tag{1.12}$$

Le problème dual (1.11) s'exprime par

$$\begin{cases} \max \frac{1}{2}x^\top Qx + c^\top x + \lambda(Ax - b) \\ Qx + A^\top\lambda = -c, \\ \lambda \geqslant 0. \end{cases} \tag{1.13}$$

De (1.12) , on a $c = -Qx - A^\top\lambda$, en remplaçant ceci dans (1.13) on obtiendra la formulation du problème dual associé au problème (1.9)

$$\begin{cases} \max -\frac{1}{2}x^\top Qx - \lambda^\top b \\ Qx + A^\top\lambda = -c \\ \lambda \geqslant 0. \end{cases}$$

En programmation quadratique, la résolution d'un programme quadratique avec contraintes mixtes passe d'abord par la résolution d'un programme quadratique avec contraintes égalités.

3.1 Programmation quadratique avec contraintes égalités

$$\begin{cases} \min f(x) = \frac{1}{2}x^\top \mathrm{Q}x + x^\top c \\ \mathrm{A}x = b. \end{cases} \tag{1.14}$$

où la matrice $\mathrm{A} \in \mathbb{R}^{m\times n}$, $(m \leq n)$, tel que A est de rang maximal $\left(rg(\mathrm{A}) = m\right)$ et $b \in \mathbb{R}^{m\times 1}$ alors les conditions nécessaires du premier ordre pour que x^* soit une solution de (1.14), est l'existence d'un vecteur λ^*, tel que le système d'équations suivant soit satisfait :

$$\begin{bmatrix} \mathrm{Q} & -\mathrm{A}^\top \\ \mathrm{A} & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix}$$

Ces conditions sont une conséquence du théorème de Lagrange. Ce système peut être réécrit sous une forme utile pour le calcul, en exprimant x^* par $x^* = x + p$ où x est une estimation de la solution et p est la direction à trouver. En introduisant cette notation et en réarrangeant les équations, nous obtenons

$$\begin{bmatrix} \mathrm{Q} & \mathrm{A}^\top \\ \mathrm{A} & 0 \end{bmatrix} \begin{bmatrix} -p \\ \lambda^* \end{bmatrix} = \begin{bmatrix} g \\ h \end{bmatrix} \tag{1.15}$$

avec $h = \mathrm{A}x - b$, $g = c + \mathrm{Q}x$ et $p = x^* - x$. La matrice (1.15) s'appelle la matrice de Karush-Kuhn-Tucker (K-K-T).

Notons par Z, la matrice de $\mathbb{R}^{n\times(n-m)}$ dont les colonnes forment une base de $\ker \mathrm{A}$. Alors, Z est de rang maximal satisfaisant $\mathrm{AZ} = 0$. Nous nommons la matrice $\mathrm{Z}^\top \mathrm{QZ}$ la matrice Hessienne réduite.

Lemme 1.1 *[30] Supposons que,* A *est une matrice de rang maximal et la matrice hessienne réduite* $\mathrm{Z}^\top \mathrm{QZ}$ *est définie positive, alors la matrice de K-K-T*

$$\mathrm{K} = \begin{bmatrix} \mathrm{Q} & \mathrm{A}^\top \\ \mathrm{A} & 0 \end{bmatrix}$$

est non-singulière, donc il existe une unique paire de vecteurs (x^*, λ^*) *vérifiant les conditions de Lagrange.*

Théorème 1.8 *[30] Supposons que* $rg(\mathrm{A}) = m$ *et que la matrice Hessienne réduite* $\mathrm{Z}^\top \mathrm{QZ}$ *est définie positive. Alors,* x^* *vérifiant le système*

$$\begin{bmatrix} \mathrm{Q} & -\mathrm{A}^\top \\ \mathrm{A} & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix} \tag{1.16}$$

est la solution unique globale du problème (1.14).

Remarque 1.1

1. *Si la matrice réduite* $\mathrm{Z}^\top \mathrm{QZ}$ *est uniquement semi définie positive, la solution* x^* *du système (1.16) est uniquement un minimum local.*
2. *Si la matrice réduite* $\mathrm{Z}^\top \mathrm{QZ}$ *est indéfinie, la solution* x^* *du système (1.16) est uniquement un point stationnaire.*

3.2 La méthode Null-Space

La méthode espace noyau, n'exige pas l'inversibilité de la matrice Q, elle exige seulement l'existence des hypothèses du lemme précédant à savoir $rg(A) = m$ et $Z^\top QZ > 0$.

Cependant, elle nécessite la connaissance de la matrice Z dont les colonnes forment une base de $\ker A$.

Rappelons que la solution vérifie le système de KKT(1.15). Supposons qu'on dispose d'une matrice $Y \in \mathbb{R}^{n\times m}$ telle que la matrice $\begin{bmatrix} Y & | & Z \end{bmatrix} \in \mathbb{R}^{n\times n}$ soit inversible. Alors tout vecteur de $p \in \mathbb{R}^n$, peut s'écrire sous la forme

$$p = Yp_Y + Zp_Z, \tag{1.17}$$

où $p_Y \in \mathbb{R}^m$ et $p_Z \in \mathbb{R}^{n-m}$ sont les coordonnées du vecteur p associées aux colonnes de Y et Z respectivement. La deuxième équation du système donne

$$Ap = -h.$$

En substituant $p = Yp_Y + Zp_Z$ et en remarquant que $AZ = 0$, on trouve $AYp_Y = -h$. Puisque A est de rang m et $\begin{bmatrix} Y & | & Z \end{bmatrix} \in \mathbb{R}^{n\times n}$ est inversible, le produit $A\begin{bmatrix} Y & | & Z \end{bmatrix} = \begin{bmatrix} AY & | & 0 \end{bmatrix}$ est de rang m. On déduit que AY est inversible et p_Y est bien déterminé. En substituant $p = Yp_Y + Zp_Z$ dans la première équation $-Qp + A^\top \lambda^* = g$, on aura

$$-QYp_Y - QZp_Z + A^\top \lambda^* = g.$$

En multipliant les deux membres par Z, on obtient

$$(Z^\top QZ)p_Z = -Z^\top QYp_Y - Z^\top g.$$

Le système $(Z^\top QZ)p_Z = -Z^\top QYp_Y - Z^\top g$ peut être résolu en effectuant une factorisation de Cholesky ([34]) de la matrice Hessienne réduite $Z^\top QZ$ pour déterminer p_Z. En suite, on peut calculer $p = Yp_Y + Zp_Z$. Pour obtenir le multiplicateur de lagrange λ^*, on multiplie la première équation $-Qp + A^\top \lambda^* = g$, par $Y^\top$, on obtient

$$(AY^\top)\lambda^* = Y^\top(g + Qp),$$

qui peut être résolu.

Remarque 1.2 *On peut utiliser la factorisation QR pour déterminer les matrices* Z *et* Y. *En effet : Comme le* $rg(A) = m$, *considérons la matrice* $A^\top \in \mathbb{R}^{n\times m}$, *le nombre de lignes* $n \geqslant m$ *le nombre de colonnes. Soient* $Q \in \mathbb{R}^{n\times n}$ *et* $R \in \mathbb{R}^{n\times m}$ *issu de la décomposition* QR *de Householder ([34]) de* $A^\top$

$$Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix},$$

où $Q_1 \in \mathbb{R}^{n\times m}$ *et* $Q_2 \in \mathbb{R}^{n\times(n-m)}$, *les colonnes des deux matrices sont orthonormales.*

On a

$$A^\top = QR \Rightarrow Q^\top A^\top = R \Rightarrow AQ = R^\top \Rightarrow \begin{bmatrix} AQ_1 & AQ_2 \end{bmatrix} = \begin{bmatrix} \widetilde{R}^\top & 0 \end{bmatrix},$$

avec $\widetilde{R}^\top \in \mathbb{R}^{m\times m}$ *et* $0 \in \mathbb{R}^{m\times(n-m)}$. *Par identification, on en déduit que* $AQ_2 = 0$ *et les colonnes de la matrices* Q_2 *forment une base de* $\ker A$. *On peut donc prendre* $Z = Q_2$ *et* $Y = Q_1$.

3.3 Gradient conjugué appliqué à la forme réduite

La méthode du gradient conjugué s'applique bien à un problème quadratique sans contraintes et elle est très éfficace quand la matrice hessienne est définie positive.

Pour le problème :

$$\begin{cases} \min f(x) = \frac{1}{2}x^\top \mathrm{Q}x - x^\top c \\ x \in \mathbb{R}^n, \end{cases} \tag{1.18}$$

la solution est caractérisée par

$$\nabla f(x^*) = 0,$$

d'où

$$x^* = \mathrm{Q}^{-1}c.$$

La méthode du gradient conjugué est conçue pour les problèmes de grande taille, car dans ce cas, inverser une matrice est trop couteûx.

Algorithme 1.1 *[12]*
1) Pour $k = 0$, *choisir un point initial* x_0, *on pose* $g^0 = \nabla f(x^0)$
2) Si $g^0 = 0$, *STOP, sinon posons* $d^0 = -g^0$.
3) Pour $k = 1, 2, ..$ *Calculons* $\alpha_k = -\frac{(g^k)^\top d^\top}{(d^k)^\top \mathrm{Q}d^k}$, $x^{k+1} = x^k + \alpha_k d^k$ *et* $g^{k+1} = \nabla f(x^{k+1})$.
4) Si $g^{k+1} = 0$ *STOP.*
5) Sinon, calculons $\beta_k = \frac{(g^{k+1})^\top \mathrm{Q}d^\top}{(d^k)^\top \mathrm{Q}d^k}$ *et* $d^{k+1} = -g^{k+1} + \beta_k d^k$. *Posons* $k = k+1$ *et allons à l'étape3.*

Reprennant le problème (1.14), on suppose qu'on dispose ici aussi des deux matrices Y et Z, comme dans (1.17). La solution optimale x^* peut s'éxprimer par

$$x^* = \mathrm{Y}x_\mathrm{Y} + \mathrm{Z}x_\mathrm{Z},$$

avec $x_\mathrm{Z} \in \mathbb{R}^{n-m}$ et $x_\mathrm{Y} \in \mathbb{R}^m$.

De la contrainte $\mathrm{A}x = b$, on aura

$$\mathrm{AY}x_\mathrm{Y} = b.$$

On substition x^* dans le problème (1.14), on trouve

$$\begin{cases} \min \frac{1}{2}x_\mathrm{Z}^\top \mathrm{Z}^\top \mathrm{QZ}x_\mathrm{Z} + x_\mathrm{Z}^\top c_\mathrm{Z} \\ x_\mathrm{Z} \in \mathbb{R}^{n-m}, \end{cases}$$

où $c_\mathrm{Z} = \mathrm{Z}^\top \mathrm{QY}x_\mathrm{Y} + \mathrm{Z}^\top c$. Ce problème peut se résoudre avec la méthode du gradient conjugué donnée par (Algorithme1.1)

3.4 Programmation Quadratique avec Contraintes Mixtes

Considérons le problème de programmation quadratique avec contraintes égalités et inégalités (contraintes mixtes) suivant :

$$\begin{cases} \min \frac{1}{2}x^\top \mathrm{Q}x + x^\top c \\ (a^i)^\top x = b_i \ , \ i \in \mathrm{I} \\ (a^i)^\top x \geqslant b_i \ , \ i \in \mathrm{J}. \end{cases} \tag{1.19}$$

Toute solution optimale x^* vérifie les conditions de K-K-T pour certain multiplicateur λ^*

$$\begin{cases} Qx^* + c - \sum_{i \in I \cup J} \lambda_i^* a^i = 0 \\ \lambda_i^* ((a^i)^\top x^* - b_i) = 0 \ i \in J \\ \lambda_i^* \geqslant 0 \ i \in J \\ (a^i)^\top x^* - b_i = 0 \ i \in I \\ (a^i)^\top x^* - b_i \geqslant 0 \ i \in J \end{cases}$$

Exploitons à présent l'ensemble

$$\mathscr{A}(x^*) = I \cup \left\{ i \in J : (a^i)^\top x^* = b_i \right\}.$$

Le système devient

$$\begin{cases} Qx^* + c - \sum_{i \in \mathscr{A}(x^*)} \lambda_i^* a^i = 0 \\ (a^i)^\top x^* = b_i \ , \ i \in \mathscr{A}(x^*) \\ (a^i)^\top x^* \geqslant b_i \ , \ i \in J \backslash \mathscr{A}(x^*) \\ \lambda_i^* \geqslant 0 \ , \ i \in J \cap \mathscr{A}(x^*) \end{cases}$$

Cas convexe

Pour un problème QP convexe, les conditions de K-K-T sont suffisantes pour que x^* soit le minimum global du problème. Ceci étant une conséquence directe du théorème (1.7).

D'où, si x^* satisfait les conditions de K-K-T

$$\begin{cases} Qx^* + c - \sum_{i \in \mathscr{A}(x^*)} \lambda_i^* a^i = 0 \\ (a^i)^\top x^* = b_i \ , \ i \in \mathscr{A}(x^*) \\ (a^i)^\top x^* \geqslant b_i \ , \ i \in J \backslash \mathscr{A}(x^*) \\ \lambda_i^* \geqslant 0 \ , \ i \in J \cap \mathscr{A}(x^*) \end{cases} \tag{1.20}$$

pour un certains λ_i^* avec $i \in \mathscr{A}(x^*)$ et Q est semi-définie positive, alors x^* est un minimum global du Problème (1.19)

$$\begin{cases} \min \frac{1}{2} x^\top Q x + x^\top c \\ (a^i)^\top x = b_i \ , \ i \in \mathscr{A}(x^*) \\ (a^i)^\top x \geqslant b_i \ , \ i \in J \backslash \mathscr{A}(x^*) \end{cases}$$

3.5 La méthode Active Set

La méthode Active-Set [30] est conçue pour résoudre un problème quadratique avec contraintes mixtes. On considère uniquement le cas où la matrice
Q est définie positive. Si l'ensemble des contraintes actives à la solution optimale $\mathscr{A}(x^*)$ est connu d'avance, on peut trouver x^*, en résolvant par les techniques déjà vus aux sections précédentes le problème

$$\begin{cases} \min q(x) = \frac{1}{2} x^\top Q x + x^\top c, \\ (a^i)^\top x = b_i \ , \ i \in \mathscr{A}(x^*). \end{cases}$$

En effet, les conditions de Lagrange pour ce problème donnent

$$\begin{cases} Qx^* + c - \sum_{i \in \mathscr{A}(x^*)} \lambda_i^* a^i = 0 \\ (a^i)^\top x^* = b_i \ , \ i \in \mathscr{A}(x^*). \end{cases}$$

il suffirait donc de prouver que les λ_i^* soient positives pour les $i \in J \cap \mathscr{A}(x^*)$. Bien entendu, on ne connait pas au par avance $\mathscr{A}(x^*)$. La détermination de cet ensemble est le principal défi auquel font face les algorithmes pour les QP avec contraintes inégalités.

Etant à l'itération k et ayant l'itéré x^k, il faut construire l'ensemble $W_k \subset \mathscr{A}\left(x^k\right)$. Cet ensemble est appelé l'ensemble de travail (Working Set).

Une exigence importante que nous imposons à W_k est que la famille de vecteurs

$$\left\{a^i\right\}_{i \in W_k} \text{ soient linéairement indépendants.}$$

On vérifie d'abord :

si x^k minimise la fonction objectif q sur l'ensemble de travail W_k.
Sinon, on résout un sous-problème quadratique avec contraintes égalité.

Le sous problème de la méthode Active Set

Soit

$$p = x - x^k \text{ et } g_k = \nabla f\left(x^k\right) = Qx^k + c$$

Comme x^k est connue, trouver p revient à trouver x. En remplaçant x dans notre fonction objectif f, on obtient

$$\begin{aligned} f(x) &= f\left(x^k + p\right) = \frac{1}{2}\left(x^k + p\right)^\top Q\left(x^k + p\right) + \left(x^k + p\right)^\top c \\ &= \frac{1}{2}(x^k)^\top Q x^k + \frac{1}{2}p^\top Q p + p^\top Q x^k + (x^k)^\top c + p^\top c \\ &= \frac{1}{2}p^\top Q p + p^\top g^k + \rho^k \end{aligned}$$

où,

$$\rho^k = \frac{1}{2}(x^k)^\top Q x^k + (x^k)^\top c.$$

Le sous problème est

$$(\mathscr{P}_k)\begin{cases} \min \frac{1}{2}p^\top Q p + p^\top g^k + \rho^k \\ (a^i)^\top p = 0\ ,\ i \in W_k. \end{cases}$$

Comme ρ^k est constant par rapport à p, le sous problème considéré à chaque itération est

$$(\mathscr{P}_k)\begin{cases} \min \frac{1}{2}p^\top Q p + p^\top g^k \\ (a^i)^\top p = 0\ ,\ i \in W_k. \end{cases} \tag{1.21}$$

Dénotons par p^k la solution du problème $(\mathscr{P}_k)$.

On a

$$\forall\ i \in W_k,\ (a^i)^\top (x^k + \alpha p^k) = (a^i)^\top x^k = b_i.$$

Puisque les contraintes dans W_k sont stisfaites pour x^k, elles le seront aussi pour $x^k + \alpha p^k$, pour tout α.

On pose

$$x^{k+1} = x^k + \alpha_k p^k,$$

avec α_k est la plus grande valeur dans $[0,1]$ pour laquelle toutes les contraintes sont satisfaites.

$$i \notin \mathrm{W}_k \text{ on a } \begin{cases} \text{soit } (a^i)^\top p^k \geqslant 0 & \text{alors } (a^i)^\top (x^k + \alpha_k p^k) \geqslant (a^i)^\top x^k \geqslant b_i \\ \text{soit } (a^i)^\top p^k < 0 & \text{alors } (a^i)^\top (x^k + \alpha_k p^k) \geqslant b_i \\ \text{que si} & \alpha_k \leqslant (b_i - (a^i)^\top x^k)/(a^i)^\top p^k \end{cases}$$

Donc, on prend

$$\alpha_k = \min\left\{1, \min_{i \notin \mathrm{W}_k, (a^i)^\top p^k < 0} \frac{b_i - (a^i)^\top x^k}{(a^i)^\top p^k}\right\}. \tag{1.22}$$

Définition 1.4 *Une contrainte i est dite bloquante si celle-ci réalise le minimum dans la relation (1.22).*

Si $\alpha_k = 1$ et aucune nouvelle contrainte n'est active au point $x^k + \alpha_k p^k$, alors il n'y a pas de contrainte bloquante sur cette itération (comme on peut avoir une contrainte bloquante avec $\alpha_k = 1$) .

On peut avoir $\alpha_k = 0$, car on peut avoir $(a^i)^\top p^k < 0$ pour une i active en x^k mais pas encore dans W_k.

Si $\alpha_k < 1$ alors W_k sera mis à jour

$$\mathrm{W}_{k+1} = \mathrm{W}_k \cup \{i \text{ la contrainte bloquante}\}.$$

On continue à itérer de cette manière, en ajoutant des contraintes à l'ensemble de travail jusqu'à ce qu'on atteint un point $\widehat{x}$ qui minimise la fonction objectif quadratique sur son ensemble de travail courant $\widehat{\mathrm{W}}$.

Il est simple de reconnaître un tel point, car le sous problème admet la solution nulle $p^k = 0$.

Si $p^k = 0$, satisfait la première condition nécessaire d'optimalité, on aura

$$\sum_{i \in \widehat{\mathrm{W}}} a^i \widehat{\lambda}_i = g = \mathrm{Q}\widehat{x} + c, \tag{1.23}$$

en affectant la valeurs 0 pour les $\widehat{\lambda}_i$ tel que $i \notin \widehat{\mathrm{W}}$. En raison du contrôle imposé sur le pas α_k, $\widehat{x}$ est également admissible par rapport à toutes les contraintes, de sorte que les deuxième et troisième conditions K-K-T données dans la relation (1.20) soient satisfaites.

Reste à tester les $\widehat{\lambda}_i$

$$\begin{cases} \text{Si tous les } \widehat{\lambda}_i \text{ sont positifs, Stop, la solution optimale } x^* = x^k, \\ \text{Sinon, mettre à jour } \mathrm{W}_{k+1} = \mathrm{W}_k - \left\{ j = \arg\min_{i \in \widehat{\mathrm{W}} \cap \mathrm{I}} \widehat{\lambda}_i \right\}. \end{cases}$$

Algorithme 1.2 *[30]*
Initialisation x^0 et W_0 (un sous ensemble des contraintes actives au point x^0).
Pour $k = 0, 1, 2, \ldots$ Résoudre le problème (1.21).
Si $p^k = 0$
Calculer les multiplicateurs de Langrange $\widehat{\lambda}_i$ avec (1.23) avec $\widehat{\mathrm{W}} = \mathrm{W}_k$.
Si $\widehat{\lambda}_i \geqslant 0$, $\forall i \in \mathrm{W}_k \cap \mathrm{I}$ alors, STOP $x^* = x^k$.
Sinon $j : \arg\min_{i \in \mathrm{W}_k \cap \mathrm{I}} \widehat{\lambda}_i$, $x^{k+1} = x^k$, $\mathrm{W}_{k+1} = \mathrm{W}_k \setminus \{j\}$.
Sinon ($p^k \neq 0$ *)*
calculer α_k avec (1.22) , $x^{k+1} = x^k + \alpha_k p^k$.
S'il existe une contrainte bloquante $\mathrm{W}_{k+1} = \mathrm{W}_k \cup \{$indice de la contrainte bloquante$\}$.
Sinon $\mathrm{W}_{k+1} = \mathrm{W}_k$.
Fin pour.

Exemple 1.2 *[30] Considérons le problème*

$$\begin{cases} \min q(x) = (x_1-1)^2 + (x_2-2.5)^2 \\ x_1 - 2x_2 + 2 \geqslant 0, \\ -x_1 - 2x_2 + 6 \geqslant 0, \\ -x_1 + 2x_2 + 2 \geqslant 0, \\ x_1 \geqslant 0, \\ x_2 \geqslant 0. \end{cases}$$

On applique la méthode Active Set, avec $x^0 = (2,0)^\top$. *On détermine les données*

$$\mathrm{Q} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},\ c = \begin{bmatrix} -2 \\ -5 \end{bmatrix},\ \mathrm{A} = \begin{bmatrix} 1 & -2 \\ -1 & -2 \\ -1 & 2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},\ b = \begin{bmatrix} -2 \\ -6 \\ -2 \\ 0 \\ 0 \end{bmatrix}. \tag{1.24}$$

Pour $x^0 = (2,0)^\top$, *on a* $a^3 = (-1,2)^\top$ *et* $a^5 = (0,1)^\top$ *sont linéairements indépendants donc* $\mathrm{W}_0 \{3,5\}$.

On calcule $g^0 = \nabla q(x^0) = (2,-5)^\top$ *et on considère le sous problème*

$$(\mathscr{P}_0) \begin{cases} \min \frac{1}{2} p^\top \mathrm{Q} p + p^\top g^0 \\ (a^i)^\top p = 0, \qquad i \in \mathrm{W}_0. \end{cases}$$

On doit résoudre le système

$$\begin{bmatrix} \mathrm{Q} & -\mathrm{A}_{\mathrm{W}_0}^\top \\ \mathrm{A}_{\mathrm{W}_0}^\top & 0 \end{bmatrix} \begin{bmatrix} p \\ \lambda \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \\ 0 \end{bmatrix}.$$

Le système est bien

$$\begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & -2 & -1 \\ -1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \widetilde{\lambda}_3 \\ \widetilde{\lambda}_5 \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \\ 0 \end{bmatrix}.$$

On trouve $p^0 = (0,0)$. *On doit trouver* $\widetilde{\lambda}_3$ *et* $\widetilde{\lambda}_5$ *en résolvant le système*

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix} \widetilde{\lambda}_3 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \widetilde{\lambda}_5 = \begin{bmatrix} 2 \\ -5 \end{bmatrix},$$

on trouve que $(\widetilde{\lambda}_3, \widetilde{\lambda}_5) = (-2,-1)$.
Mise à jour de W

$$\mathrm{W}_1 = \{5\} \quad et \quad x^1 = x^0.$$

Itération 1
Comme $x^1 = x^0$ *et* $g^1 = g^0$, *on doit résoudre le problème*

$$(\mathscr{P}_1) \begin{cases} \min \frac{1}{2} p^\top \mathrm{Q} p + p^\top g^1 \\ (a^i)^\top p = 0, \qquad i \in \mathrm{W}_1. \end{cases}$$

On résout le sous problème $(\mathscr{P}_1)$

$$\begin{bmatrix} Q & -A_{W_1}^\top \\ A_{W_1}^\top & 0 \end{bmatrix} \begin{bmatrix} p \\ \lambda \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \\ 0 \end{bmatrix}.$$

Explicitement, on a le système

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \widetilde{\lambda}_5 \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \end{bmatrix}.$$

On touve que $p^1 = (-1,0)^\top \neq 0$. *On calcule le pas* α_1 *tel que*

$$\alpha_1 = \min\left\{1, \min_{i \notin W_1, (a^i)^\top p^1 < 0} \frac{b_i - (a^i)^\top x^1}{(a^i)^\top p^1}\right\} = \min\{1,4,4\} = 1.$$

Il n'y a pas de contraintes bolquante, et le nouveau itéré est

$$x^2 = x^1 + \alpha_1 p^1 = (1,0)^\top \text{ et } W_2 = W_1.$$

Itération 2

On calcule $g^2 = Qx^2 + c = (0,-5)^\top$. *La résolution du nouveau sous problème* $(\mathscr{P}_2)$ *donne* $p^2 = (0,0)^\top$ *et* $\widetilde{\lambda}_5 = -5$ *et donc* $W_3 = \varnothing$ *et* $x^3 = x^2$.

Itération 3

Oncalcule $g^3 = Qx^3 + c = (0,-5)^\top$. *La résolution du nouveau sous problème* $(\mathscr{P}_3)$, *donne* $p^3 = (0,2.5)^\top$, $\alpha_3 = 0.6$ *(contrainte bloquante) et donc* $W_4 = \{1\}$ *et aussi* $x^4 = (1,1.5)^\top$.

Itération 4

On calcule $g^4 = Qx^4 + c = (0,-2)^\top$. *La résolution du nouveau sous problème* $(\mathscr{P}_4)$, *donne* $p^4 = (0.4,0.2)^\top$ *et* $\widetilde{\lambda}_1 = 0.8$, $\alpha_4 = 1$ *et donc* $x^5 = (1.4,1.7)^\top$.

Itération 5

On calcule $g^5 = Qx^5 + c = (0.8,-1.6)^\top$. *La résolution du nouveau sous problème* $(\mathscr{P}_5)$, *donne* $p^5 = (0,0)^\top$ *et* $\widetilde{\lambda}_1 = 0.8$, *et donc* $x^* = x^5 = (1.4,1.7)^\top$ *solution optimale.*

Chapitre 2

Les Machines à Vecteurs de Support

1 Introduction

Les machines à vecteurs de support (SVMs) sont des algorithmes ayant comme but de résoudre les problèmes de classification et de régression. On s'intéresse dans ce travail uniquement au problème de classification binaire qui est un problème de classification à deux classes, dans lequel on tente de déterminer la classe à laquelle appartient un individu (individu est ici employé au sens de constituant d'un ensemble de données) parmi deux choix possibles. Pour ce faire, on utilise les caractéristiques connues de cet individu. Ces n caractéristiques sont représentées par un vecteur $x \in \mathbb{R}^n$. La classe à laquelle appartient l'individu est représentée par $y \in \{-1,1\}$. Les machines à vecteurs de support utilisent un ensemble de données pour lesquelles le classement est déjà connu et s'en servent pour construire une règle qui permet d'effectuer une bonne classification. Cet ensemble de données est appelé l'ensemble d'apprentissage. La règle trouvée avec l'ensemble d'apprentissage doit être la plus générale possible, puisqu'il faut aussi qu'elle soit bonne pour de nouvelles données qui n'étaient pas dans l'ensemble d'apprentissage. Nous présentons ici comment les SVMs font pour trouver cette règle d'abord dans le cas le plus simple possible, c'est-à-dire le cas où les données sont linéairement séparables.

2 Hyperplan séparateur

Supposons que nous disposons d'un ensemble d'apprentissage de m données de la forme $(x^i, y_i) \in \mathbb{R}^n \times \{-1,1\}$ $(i = 1,...,m)$, dont nous voulons nous servir pour déterminer une règle permettant de classer les données. Supposons aussi que ces données sont linéairement séparables, c-à-d qu'il existe un hyperplan dans $\mathbb{R}^n$ tel que toutes les données appartenant à la classe 1 (P_+) se retrouvent d'un côté de l'hyperplan alors que celles de la classe -1 (P_-) se situent de l'autre côté (voir figure 2.1).

Plus formellement, les données sont dites linéairement séparables s'il existe un hyperplan

$$w^\top x + \gamma = 0,$$

tel que $w^\top x + \gamma > 0$ pour tout x appartenant à P_+, et $w^\top x + \gamma < 0$ pour tout x appartenant à P_-, avec $w = (w_1,...,w_n) \in \mathbb{R}^n$ le vecteur normal à l'hyperplan et $\gamma \in \mathbb{R}$ un scalaire appelé le biais (remarquons que tout hyperplan peut s'écrire sous cette forme). Nous dirons d'un tel hyperplan qu'il sépare les données. En effet, il suffit de prendre un hyperplan qui sépare les classes, puis de classer les données selon le côté de l'hyperplan où elles se trouvent.

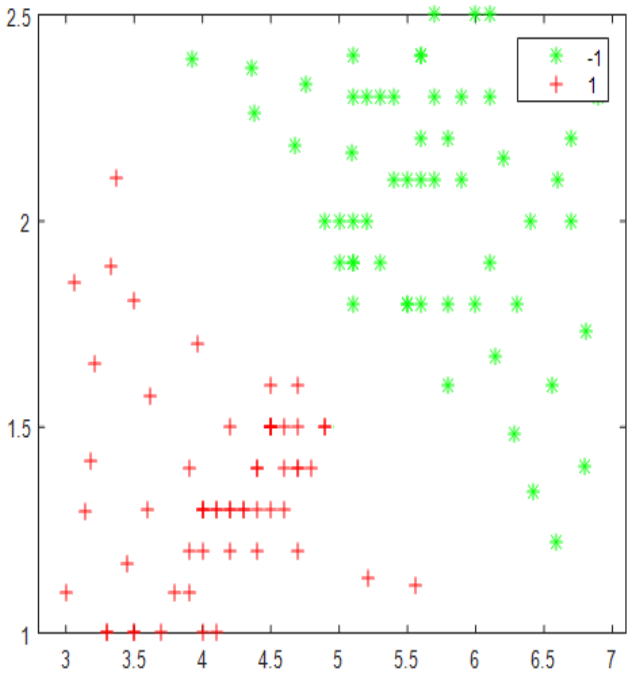


FIGURE 2.1 – Echantillon de données linéairement séparables

Plus formellement, soit

$$w^\top x + \gamma = 0$$

un hyperplan qui sépare les données. Alors, il suffit d'utiliser la fonction suivante (appelée la fonction $classificatrice$) pour effectuer la classification (voir figure 2.2) :

$$f(x) = w^\top x + \gamma. \tag{2.1}$$

$$x \in \mathrm{P}_+ \Longrightarrow f(x) \geqslant 0,$$
$$x \in \mathrm{P}_- \Longrightarrow f(x) \leqslant 0.$$

Cependant, si les données sont linéairement séparables, il existe une infinité d'hyperplans qui peuvent servir de séparateurs (voir figure 2.3). L'idée des machines à vecteurs de support est de choisir le meilleur hyperplan, c-à-d celui qui donnera la règle qui se généralisera le mieux à d'autres données que celles de l'ensemble d'apprentissage. Afin de déterminer ce qui caractérise le meilleur hyperplan, introduisons le concept de marge.

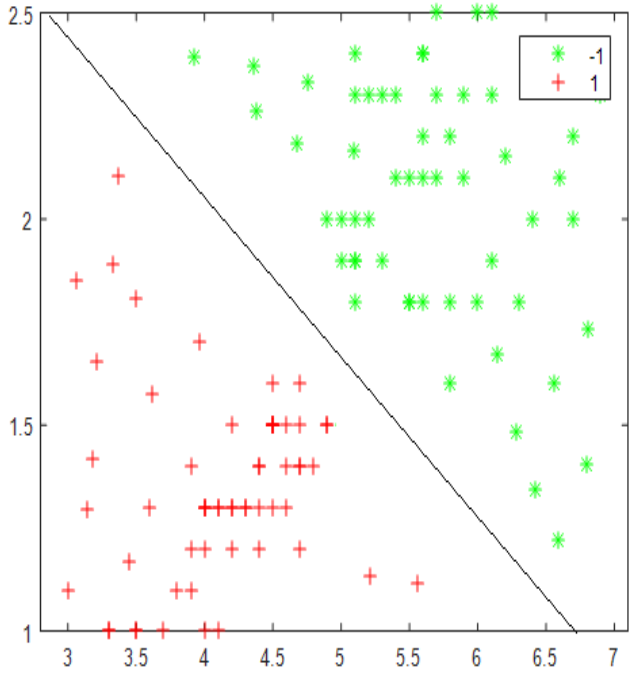


FIGURE 2.2 – hyperplan séparateur des données linéairement séparables

3 Marge et hyperplan canonique

Définissons la marge d'un hyperplan comme étant la distance entre l'hyperplan et la donnée la plus proche. Plus formellement (si on note par s la donnée de l'échantillon le plus proche et on prend la norme 2), on a

$$d(s,\mathrm{H}) = \frac{|w^\top s+\gamma|}{\|w\|_2},\ avec\ \mathrm{H} = \{x, w^\top x+\gamma = 0\}.$$

Les travaux de Vapnik sur l'apprentissage statistique [44] et [45], ont montré que l'hyperplan qui aura la meilleure généralisation est celui qui possède la plus grande marge. Ce concept est à la base des machines à vecteurs de support. Dans le cas où les données sont linéairement séparables, les SVMs trouvent l'hyperplan qui sépare les données avec la plus vaste marge possible, puis utilisent cet hyperplan pour classer de nouvelles données à l'aide de la fonction classificatrice (2.1). Toutefois, le problème de trouver l'hyperplan avec la marge maximale est mal posé, puisqu'il existe en réalité une infinité de manières différentes d'écrire le même hyperplan. En effet, supposons que l'hyperplan

$$w^\top x+\gamma = 0$$

soit un hyperplan dont la marge est maximale, et soit $\lambda \in \mathbb{R}^+\setminus\{0\}$. Alors, l'hyperplan

$$\lambda w^\top x+\lambda\gamma = 0$$

est en réalité le même hyperplan et sépare les données, puisque λ est positif. Par conséquent, $\lambda w^\top x+\lambda\gamma = 0$ correspond aussi à l'hyperplan dont la marge est maximale, mais possède un vecteur des coefficients et un biais différents (si $\lambda \neq 1$). Le nombre infini de manières d'écrire la solution du problème de l'hyperplan avec la plus vaste marge complique sa résolution. Afin de rendre le problème bien posé et pour lutter contre une

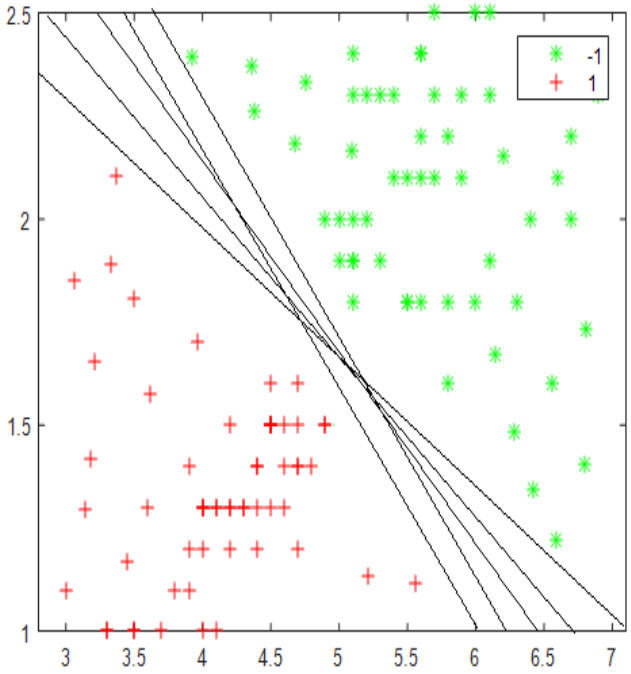


FIGURE 2.3 – Plusieurs hyperplans séparateurs

solution insignifiante (car $w = 0_{\mathbb{R}^m}$ et $\gamma = 0$ vérifient l'équation de l'hyperplan), introduisons le concept d'hyperplan canonique (voir figure 2.4). Pour cela, on normalise les deux hyperplans de bondissements (voir figure 2.4), comme suit :

$$x \in \mathrm{P}_+ \Longrightarrow f(x) = w^\top x + \gamma \geqslant 1,$$
$$x \in \mathrm{P}_- \Longrightarrow f(x) = w^\top x + \gamma \leqslant -1.$$

Si s appartient à l'hyperplan de bondissement supérieur, alors

$$s \in \mathrm{P}_+ \Longrightarrow f(s) = w^\top s + \gamma = 1,$$

si s appartient à l'hyperplan de bondissement inférieur, alors

$$s \in \mathrm{P}_- \Longrightarrow f(s) = w^\top s + \gamma = -1.$$

On peut aussi montrer que tout hyperplan qui sépare les données peut s'écrire sous une forme canonique et qu'il n'existe qu'une seule façon d'écrire un hyperplan pour qu'il soit canonique. Ainsi, en ne considérant que les hyperplans canoniques, chaque hyperplan s'écrit de manière unique. De plus, il n'existe qu'un seul hyperplan pour lequel la marge est maximale. Ceci deviendra évident un peu plus loin, puisque le vecteur des coefficients de l'hyperplan sera exprimé comme étant le point qui minimise une fonction strictement convexe (rappelons que les fonctions strictement convexes n'ont qu'un unique minimum global). Par conséquent, en ne considérant que les hyperplans canoniques, le problème de trouver l'hyperplan avec la plus grande marge est bien posé.

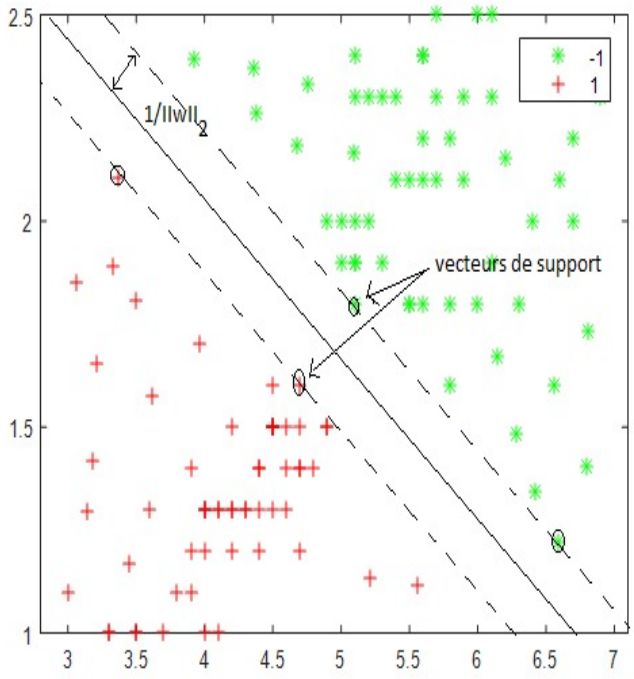


FIGURE 2.4 – Hyperplan de séparation et hyperplans de bondissement

4 Trouver l'hyperplan

On peut montrer que pour un hyperplan canonique, la marge est donnée par l'expression

$$\frac{1}{\|w\|_2},$$

on voit donc que plus $\|w\|_2$ est petite, plus la marge de l'hyperplan canonique correspondant est grande. Ainsi, afin de trouver l'hyperplan qui sépare le mieux les données, il faut trouver celui qui respecte les conditions d'un hyperplan canonique et pour lequel $\|w\|_2$ est minimale. La recherche du meilleur hyperplan peut donc s'écrire sous la forme du problème d'optimisation suivant :

$$\begin{cases} \min\limits_{w,\gamma} \quad \|w\|_2 \\ \mathrm{D}(\mathrm{A}w + e\gamma) \geqslant e, \end{cases}$$

où A est une matrice de taille $m \times n$, qui assemble tous les points x^i $(i = 1, ..., m)$ de l'échantillon

$$\mathrm{A} = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^m \end{pmatrix} = \begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_n^1 \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \cdots & \vdots \\ x_1^m & x_2^m & \cdots & x_n^m \end{pmatrix},$$

D est une matrice diagonale

$$\mathrm{D}_{ii} = y_i,$$

et

$$e = (1, \cdots, 1)^\top.$$

Nous avons ainsi formulé un problème d'optimisation dont la solution optimale est l'hyperplan canonique séparant les données avec la plus vaste marge possible. Cependant, il est possible de formuler un problème équivalent, mais avec une fonction objectif plus simple. En effet, comme

$$\|w\|_2 = \sqrt{w^\top w},$$

minimiser $\|w\|_2$ est équivalent à minimiser $\|w\|_2^2$. Évidemment, minimiser $w^\top w$ est équivalent à minimiser $\frac{1}{2} w^\top w$ (cette petite modification a été introduite afin de simplifier les calculs).

Par conséquent, afin de trouver l'hyperplan canonique qui sépare les données avec la plus grande marge possible, il suffit de résoudre le problème d'optimisation suivant :

$$\begin{cases} \min\limits_{w,\gamma} & \frac{1}{2} w^\top w \\ & \mathrm{D}(\mathrm{A}w + e\gamma) \geqslant e, \end{cases}$$

qui est un problème d'optimisation avec une fonction objectif quadratique strictement convexe. Ceci assure qu'il n'y a pas de minimum relatif et qu'il n'existe qu'une unique solution optimale.

5 Les vecteurs de support

Comme la fonction $w \longmapsto w^\top w$ est une fonction convexe continue et dérivable, les contraintes $\mathrm{D}(\mathrm{A}w + e\gamma) \geqslant e$ sont des fonctions affines et le domaine du problème est $\mathbb{R}^n$, la solution optimale trouvée respecte nécessairement les conditions de K-K-T. En particulier, elle respecte la condition de complémentarité de K-K-T, c-à-d que

$$\widetilde{\alpha}_i (\mathrm{D}_{ii}(\mathrm{A}_{i.}\overline{w} + \overline{\gamma}) - 1) = 0, \quad i = 1, \cdots, m,$$

où $\widetilde{\alpha}$ représente la solution optimale du problème dual et $(\overline{w}, \overline{\gamma})$ représente celle du problème primal. Cette condition implique que si $\mathrm{D}_{ii}(\mathrm{A}_{i.}\overline{w} + \overline{\gamma}) - 1 \neq 0$, alors $\alpha_i = 0$. Par conséquent, les seuls cas où α_i peut ne pas être nul sont ceux où $\mathrm{D}_{ii}(\mathrm{A}_{i.}\overline{w} + \overline{\gamma}) - 1 = 0$, c-à-d ceux où

$$\mathrm{D}_{ii}(\mathrm{A}_{i.}\overline{w} + \overline{\gamma}) = 1.$$

Or, les seuls points où $\mathrm{D}_{ii}(\mathrm{A}_{i.}\overline{w} + \overline{\gamma}) = 1$ sont ceux qui sont sur la marge. Par conséquent, seuls les points sur la marge peuvent avoir des α_i non nuls. Ces points sont appelés les vecteurs de support. La raison de ce nom est que ce sont les seuls points utiles pour déterminer l'hyperplan. Ainsi, tout point qui n'est pas sur la marge n'apporte aucune contribution, puisque α_i est alors nul. Si tous les points sauf les vecteurs de support étaient retirés de l'ensemble d'apprentissage, on retrouverait le même hyperplan. Les vecteurs de support peuvent donc être vus comme les points contenant toute l'information essentielle du problème.

6 Marges souples

6.1 Machines à vecteurs de support et bruit

En pratique, les données sont rarement parfaites. Il y a souvent du bruit, c-à-d des données qui sont mal classées par un modèle qui est toutefois excellent en général (voir

figure 2.5). Il s'agit donc des erreurs qui sont inévitables, même pour les meilleurs modèles. Toutefois, l'approche utilisée précédemment ne permet pas de tenir compte de ce phénomène, puisque dans les contraintes, toutes les données doivent être correctement classées. Supposons par exemple qu'un ensemble de données serait très bien séparé par un hyperplan, mais qu'il n'est pas linéairement séparable dû à la présence d'un certain bruit dans les données. Dans un tel cas, il serait impossible de construire une SVM linéaire, car il est impossible que toutes les contraintes soient respectées.

6.2 Marge souple

Un meilleur moyen serait de permettre à quelques données d'être à l'intérieur de la marge ou du mauvais côté de l'hyperplan. Il s'agit du concept de marge souple (soft margin). Une première idée serait de tenter de maximiser la marge tout en minimisant le nombre de données mal classées. Toutefois, le nombre de données mal classées peut être trompeur, puisqu'il ne permet pas de déterminer si une donnée était presque correctement classée ou si elle était en réalité très loin de l'hyperplan. Une meilleure idée est d'attribuer à chaque donnée x^i une valeur z_i qui représente à quel point la donnée est éloignée d'un bon classement, puis de tenter de minimiser la somme des z_i. Plus formellement, au lieu d'imposer

$$\mathrm{D}(\mathrm{A}w + e\gamma) \geqslant e,$$

ce qui oblige les données à être bien classées, les contraintes seront plutôt

$$\mathrm{D}(\mathrm{A}w + e\gamma) + z \geqslant e, \quad z \geqslant 0.$$

où $z = (z_1, z_2, ..., z_m)^\top$, z_i est appelées variables artificielles (slack variable).

Par conséquent, il est possible pour une donnée d'être du mauvais côté de la marge, si z_i est non nul. On dira d'une donnée qu'elle est du mauvais côté de la marge si elle est mal classée ou si sa distance par rapport à l'hyperplan séparateur est plus petite que la marge (remarquons que les points pour lesquels $z_i \neq 0$ ne sont pas considérés dans le calcul de la marge). L'objectif est ainsi de maximiser la marge tout en minimisant la somme des z_i. Le problème d'optimisation devient alors

$$\begin{cases} \min\limits_{w,\gamma,z} \ \frac{1}{2} w^\top w + \mathrm{C} e^\top z \\ \mathrm{D}(\mathrm{A}w + e\gamma) + z \geqslant e, \\ z \geqslant 0. \end{cases} \tag{2.2}$$

où $\mathrm{C} > 0$ est une constante qui représente la pénalité d'avoir des données mal classées. Lorsque C est très élevée, il y aura très peu de données mal classées, alors qu'il y en aura plus pour une valeur plus faible de cette constante. Le choix de C a une grande influence sur le modèle. En pratique, plusieurs modèles sont souvent construits, avec différentes valeurs de C, puis le meilleur est choisi. Notons ici, que si on dévelope la contraintes matricielle dans (2.2), ce dernier se formule aussi comme suit

$$\begin{cases} \min \frac{1}{2} w^\top w + \mathrm{C} \sum\limits_{i=1}^{m} z_i \\ y_i(w^\top x^i + \gamma) \geqslant 1 - z_i, \quad i = 1, \cdots, m \\ z_i \geqslant 0. \end{cases}$$

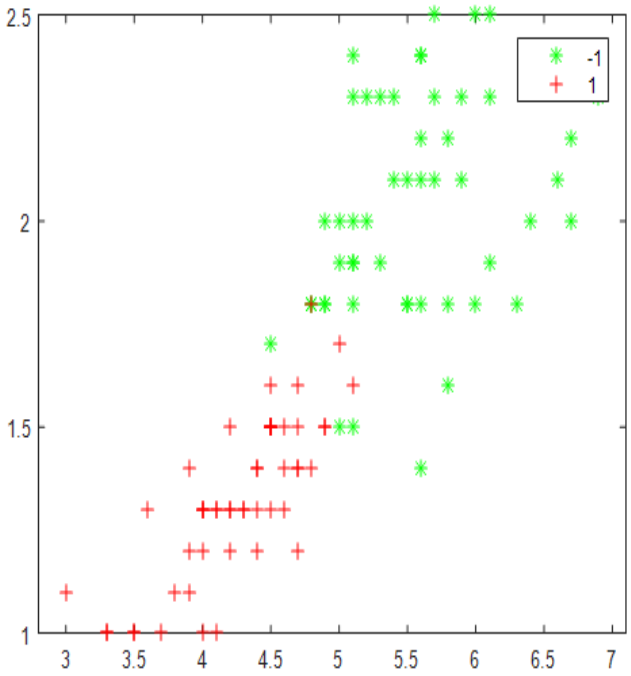


FIGURE 2.5 – Fisheriris data (données non linéairement séparables)

7 Représentation duale

Nottons par α et β les multiplicateurs de K-K-T pour les contraintes $D(Aw+e\gamma)+z \geqslant e$ et $z \geqslant 0$, la fonction Lagrangienne et les conditions de K-K-T pour le problème (2.2) peuvent être écrites comme suit :

$$\mathscr{L}(w,\gamma,z,\alpha,\beta) = \frac{1}{2}w^\top w + C\sum_{i=1}^{m} z_i - \sum_{i=1}^{m}\alpha_i(y_i(w^\top x^i+\gamma)-1+z_i) - \sum_{i=1}^{m}\beta_i z_i.$$

Les conditions d'optimalités donnent

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial w} &= 0 \Longrightarrow w = \sum_{i=1}^{m}\alpha_i y_i x^i,\\
\frac{\partial \mathscr{L}}{\partial z_i} &= 0 \Longrightarrow C = \alpha_i + \beta_i,\ i = 1,\ldots,m,\\
\frac{\partial \mathscr{L}}{\partial \gamma} &= 0 \Longrightarrow \sum_{i=1}^{m}\alpha_i y_i = 0.
\end{aligned}$$

Ces équations sont utilisées pour éliminer les variables w, z, γ et β de la fonction Lagrangienne :

$$\begin{cases}
\max\limits_{\alpha} \frac{-1}{2}\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} y_i y_j \alpha_i \alpha_j (x^i)^\top x^j + \sum\limits_{j=1}^{m}\alpha_i\\
0 \leqslant \alpha_i \leqslant C,\\
\sum\limits_{i=1}^{m} y_i \alpha_i = 0.
\end{cases} \tag{2.3}$$

Il n'est pas difficile de voir que le problème (2.3) est équivalent à :

$$\begin{cases} \min\limits_{\alpha} \frac{1}{2}\alpha^\top \mathrm{DAA}^\top \mathrm{D}\alpha - e^\top \alpha \\ 0 \leqslant \alpha \leqslant \mathrm{C}e, \\ e^\top \mathrm{D}\alpha = 0. \end{cases}$$

En effet, il est souvent plus commode de résoudre cette forme du problème pour la variable duale α et puis de récupérer w de la 1 ére condition de K-K-T, on trouve $w = \mathrm{A}^\top \mathrm{D}\alpha$ et γ en tant que multiplicateur de Lagrange pour la contrainte $e^\top \mathrm{D}\alpha$.

Remarque 2.1 *Les conditions de K-K-T tiennent toujours dans le cas de la marge souple. Par conséquent, d'après la condition de complémentarité, pour la solution optimale, les égalités suivantes sont vérifiées :*

$$\alpha^\top (\mathrm{D}(\mathrm{A}w + e\gamma) + z - e) = 0 \Leftrightarrow \alpha_i (\mathrm{D}_{ii}(\mathrm{A}_{i.}w + \gamma) + z_i - 1) = 0, \;\; i = 1, \cdots, , m,$$

$$(\mathrm{C}e - \alpha)^\top z = 0 \Leftrightarrow (\mathrm{C} - \alpha_i) z_i = 0, \; i = 1, \cdots, m.$$

Ceci implique que si $z_i \neq 0$, *alors* $\mathrm{C} - \alpha_i = 0$ *(puisque* $(\mathrm{C} - \alpha_i) z_i = 0$*), et donc* $\mathrm{C} = \alpha_i$. *De plus, si un point est tel que* $z_i \neq 0$, *alors il est du mauvais côté de la marge, ce qui découle directement du rôle de* z_i *dans le problème d'optimisation. À l'opposé, tous les points pour lesquels* $z_i = 0$ *sont du bon côté de la marge, et sont ainsi nécessairement bien classés. D'autre part, si, pour une certaine donnée, on a :* $0 < \alpha_i < \mathrm{C}$, *alors celle-ci est exactement sur la marge. En effet, on a alors* $0 < \alpha_i < \mathrm{C}$, $\alpha_i \neq \mathrm{C}$, *et donc il faut que* $z_i = 0$ *pour que* $(\mathrm{C} - \alpha_i) z_i = 0$. *De plus,* $\alpha_i \neq 0$, *ce qui implique que* $\mathrm{D}_{ii}(\mathrm{A}_{i.}w + \gamma) + z_i - 1 = 0$ *afin de respecter l'égalité* $\alpha_i (\mathrm{D}_{ii}(\mathrm{A}_{i.}w + \gamma) + z_i - 1) = 0$. *Comme* $z_i = 0$, *il s'ensuit que*

$$\mathrm{D}_{ii}(\mathrm{A}_{i.}w + \gamma) = 1,$$

et donc que $\mathrm{A}_{i.}$ *(qui représente le point* x^i*) est directement sur la marge. Enfin, les points pour lesquels* $z_i = 0$ *et* $\mathrm{D}_{ii}(\mathrm{A}_{i.}w + \gamma) - 1 \neq 0$ *ont un* α_i *nul, afin de respecter l'égalité* $\alpha_i (\mathrm{D}_{ii}(\mathrm{A}_{i.}w + \gamma) + z_i - 1) = 0$. *Les points qui sont directement sur la marge sont appelés vecteurs de support libres (free support vectors), ou encore vecteurs de support non-bornés (unbounded support vectors). Les points pour lesquels* $\alpha_i = \mathrm{C}$ *sont quant à eux appelés vecteurs de support bornés (bounded support vectors). Ici encore, les vecteurs de support sont les seuls points qui sont vraiment important pour déterminer l'hyperplan optimal, puisque ce sont les seuls points pour lesquels* $\alpha_i \neq 0$.

Il sera alors pertinent de répartir les m inconnues α_i, $i = 1, \cdots, m$ du problème en trois groupes de points $[1; m] = \mathrm{I}_0 \cup \mathrm{I}_s \cup \mathrm{I}_c$, définis en fonction de la valeur du multiplicateur de Lagrange α associé :

- $[\mathrm{I}_s]$ le groupe des points $supports$ est celui des vecteurs supports candidats, c'est-à-dire pour lesquels $0 < \alpha_i < \mathrm{C}$. Ces points sont à l'intérieur de la boîte et satisfont les contraintes. Il est aussi appelé l'ensemble de travail ($working\ set$).

- $[\mathrm{I}_c]$ le groupe des points $saturés$. Ces points sont sur le bord de la boîte. Si ces points sont trop proches de la classe opposée, voire complètement dans la classe opposée (erreur d'étiquetage par exemple), limiter leur contribution à la frontière de décision permet de régulariser la solution. Dans la solution finale ces points auront tous la valeur de leur α fixée à C : ils seront contraints et contribueront à la décision.

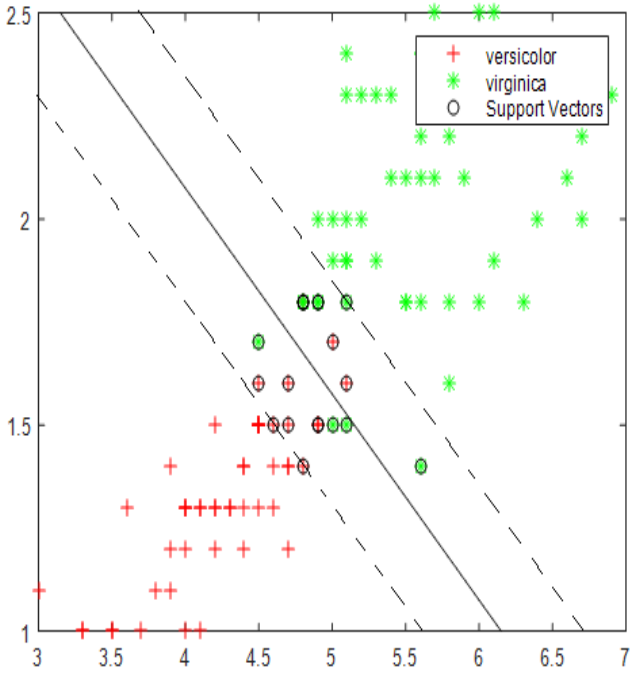


FIGURE 2.6 – Marge souple

– [I_0] le groupe des points $inactifs$. Ces points se situent également sur l'arête de la boîte. Dans ce cas, les vecteurs sont loin de la frontière entre les classes et ne contribuent en rien à la solution. Dans la solution finale ces points auront tous les valeurs de leur α fixée à 0 : ils seront contraints.

Ces trois ensembles nous aident à reformuler le problème d'optimisation :

$$\begin{cases} \max\limits_{\alpha\in\mathbb{R}^m} \frac{-1}{2}\alpha^\top DAA^\top D\alpha + e^\top\alpha \\ e^\top D\alpha = 0 \\ 0 < \alpha_i < C, \quad i \in I_s \\ \alpha_i = C, \quad i \in I_c \\ \alpha_i = 0, \quad i \in I_0. \end{cases}$$

Relation entre variables duales et primales :

Ensemble	Contraintes initiales	Contraintes primales	Contraintes Duales
I_s	$y_i(A_{i.}w+\gamma)=1$	$z_i=0$	$0<\alpha<C$
I_c	$y_i(A_{i.}w+\gamma)=1-z_i$	$z_i>0$	$\alpha=C$
I_0	$y_i(A_{i.}w+\gamma)>1$	$z_i=0$	$\alpha=0$

8 Revue de littérature des méthodes d'optimisation pour les SVMs

La formulation des SVMs représente un problème d'optimisation convexe, et bien qu'il y ait un certain nombre de méthodes d'optimisation pour résoudre ce type de problème, la formulation de la machine à vecteurs de support (SVM) est particulièrement

difficile car le nombre de contraintes sont de l'ordre de la taille de l'ensemble de données (la matrice Hessienne du problème quadratique). L'handicape des méthodes issues de la communauté optimisation est la nécessité de stocker l'intégralité de la matrice des données en mémoire et pire encore, calculer l'inverse de la matrice Hessienne à chaque itération.

À titre d'exemple, un ensemble de données contenant seulement 10 000 points de données nécessiterait 800 MBytes de RAM pour stocker la matrice Hessienne entière avec la double précision de virgule flottante. En conséquence, élaborer des méthodes rapide pour les SVMs reste un domaine de recherche très actif. Un échantillon des techniques pertinentes comprend les méthodes de points intérieurs [14], les méthodes d'ensemble actif [37], [24] gradient projeté [30], méthode de Newton-like [18], gradient conjugué projeté [48], optimisation séquentielle minimale (SMO) [33], segmentation [44], et décomposition [20].

9 Machines à vecteurs de support pour données non linéairement séparables

9.1 Transformations

Jusqu'à présent, les machines à vecteurs de support ainsi introduites, permettent de trouver une règle pour classer les données lorsque celles-ci sont linéairement séparables. Cependant, il existe bien des cas pour lesquels il est impossible de séparer entièrement les données avec un hyperplan (voir figure 2.7 et 2.8). Afin de régler ce problème, il est possible d'appliquer une transformation aux données de sorte qu'une fois transformées, elles soient linéairement séparables. L'espace où se trouvent les données avant d'être transformées est appelé l'espace d'entrée (input space), alors qu'après avoir appliqué la transformation, les données se trouvent dans ce qu'on appelle l'espace de redescription ou caractéristique (feature space). Il suffit alors de trouver l'hyperplan dans l'espace de redescription qui sépare le mieux ces données transformées. De retour dans l'espace d'entrée, le séparateur n'est pas linéaire.

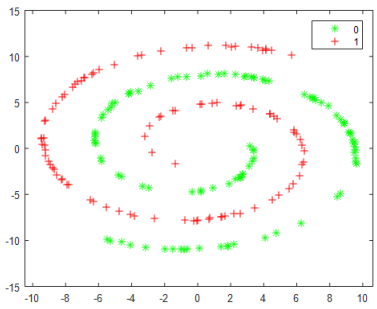


FIGURE 2.7 – Two spiral data

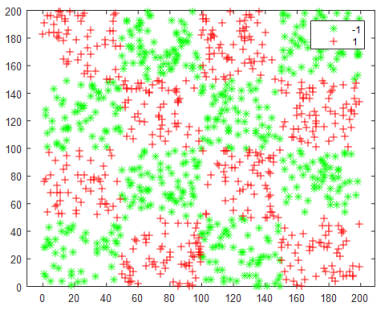


FIGURE 2.8 – 4X4 checkerboard

Soit

$$\begin{aligned} \phi : \mathbb{R}^n &\longrightarrow \mathbb{R}^{n'} \\ x &\longrightarrow \phi(x) \end{aligned}$$

la transformation appliquée aux données pour les rendre linéairement séparables, avec n' la dimension de l'espace de redescription. Très souvent, $n' > n$, ce qui signifie que

la transformation amène les données dans un espace de dimension supérieure afin de mieux pouvoir les séparer (voir figure 2.9).

Pour trouver le séparateur, on procède de la même manière que précédemment, mais en substituant $\phi(A_{i.}^{\top})$ à $A_{i.}$ $(i = 1, ..., m)$. Il s'agit donc de résoudre le problème suivant

$$\begin{cases} \min\limits_{w,\gamma,z} \ \frac{1}{2} w^{\top} w + Ce\top z \\ D_{ii}(w^{\top}\phi(A_{i.}^{\top}) + \gamma) + z_i \geqslant 1, \\ z_i \geqslant 0, \ i = 1, \cdots, m. \end{cases}$$

Le dual de ce problème est

$$\begin{cases} \min\limits_{\alpha} \frac{1}{2} \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} \alpha_i D_{ii} \phi(A_{i.}^{\top})^{\top} \phi(A_{i.}^{\top}) D_{jj} \alpha_j - \sum\limits_{i=1}^{m} \alpha_i \\ 0 \leqslant \alpha \leqslant Ce, \\ e^{\top} D\alpha = 0. \end{cases}$$

La fonction classificatrice associée à ce problème est par conséquent

$$f(x) = w^{\top}\phi(x) + \gamma = \sum_{i=1}^{m} D_{ii}\alpha_i \phi(A_{i.}^{\top})^{\top}\phi(x) + \gamma.$$

Si la transformation utilisée est appropriée, la résolution d'un de ces problèmes (le primal ou le dual) permet de trouver un séparateur non linéaire avec la marge la plus grande possible, permettant ainsi d'utiliser les machines à vecteurs de support dans le cas où les données ne peuvent pas être séparées linéairement.

9.2 Les noyaux

Toutefois, l'utilisation des transformations pose certains problèmes. En effet, il faut choisir une bonne transformation et l'appliquer à toutes les données, puis effectuer les calculs avec ces données transformées, c-à-d dans l'espace de redescription. Or, comme la dimension de cet espace est bien souvent beaucoup plus grande que celle de l'espace d'entrée, les calculs requis peuvent devenir extrêmement longs à effectuer. C'est ici que la formulation duale du problème d'optimisation prend toute son importance. En effet, on remarque que lorsque le problème est sous sa forme duale, les données de l'ensemble d'apprentissage n'apparaissent que dans un produit scalaire avec d'autres données du même ensemble. Il en est de même dans la fonction classificatrice. Ceci amène à définir comme suit une fonction appelée noyau (kernel) :

$$\begin{aligned} K : \mathbb{R}^n \times \mathbb{R}^n &\longrightarrow \mathbb{R} \\ (x^i, x^j) &\longrightarrow K(x^i, x^j) = \phi(x^i)^{\top}\phi(x^j) \end{aligned}$$

alors $K_{ij} = \phi(A_{i.}^{\top})^{\top}\phi(A_{j.}^{\top}) = K(A_{i.}^{\top}, A_{j.}^{\top})$.

Cette fonction prend en entrée deux points dans l'espace d'entrée et calcule leur produit scalaire dans l'espace de redescription. L'avantage d'une telle fonction est qu'il n'est pas nécessaire d'appliquer une transformation aux données afin de calculer leur produit scalaire dans l'espace caractéristique. Ce calcul peut se faire directement à partir des données de l'espace d'entrée.

Grâce au concept de noyau, il est possible de réécrire le problème dual de cette manière :

$$\begin{cases} \min\limits_{\alpha} \frac{1}{2} \alpha^{\top} DKD\alpha - e^{\top}\alpha \\ 0 \leqslant \alpha \leqslant Ce, \\ e^{\top} D\alpha = 0. \end{cases}$$

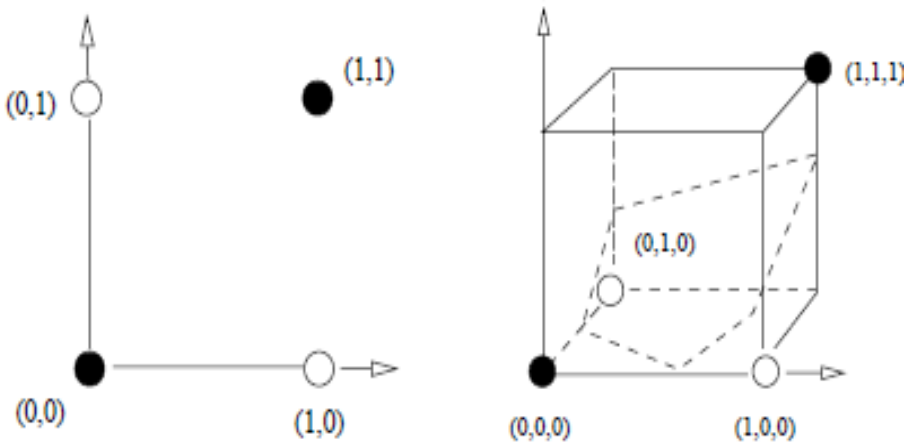


FIGURE 2.9 – transformation des données dans un autre espace

où K est la matrice $(K_{ij})_{i,j=1,\cdots,m}$,

La fonction classificatrice s'écrit ainsi :

$$f(x) = \sum_{i=1}^{m} D_{ii}\alpha_i K(A_{i.}^{\top}, x) + \gamma.$$

On remarque que de cette manière, lorsque la fonction noyau est connue, la transformation $\phi(x)$ n'apparaît nulle part, ni dans le problème, ni dans l'application de la solution. Par conséquent, grâce à la fonction noyau, il n'est pas nécessaire d'effectuer la transformation sur les données. Cette fonction permet donc de faire tous les calculs nécessaires sans avoir à se préoccuper de la dimension de l'espace caractéristique.

10 Exemples de noyaux

Il est bien de savoir qu'un noyau est tout ce qui est nécessaire pour utiliser les SVMs dans le cas non linéaire. Les noyaux les plus fréquemment utilisés pour les machines à vecteurs de support sont :

– Noyau linéaire

$$K(x^i, x^j) = (x^i)^{\top} x^j.$$

– Noyau polynomial de degré d

$$K(x^i, x^j) = ((x^i)^{\top} x^j + 1)^d.$$

– Noyau Gaussien

$$K(x^i, x^j) = \exp(-\frac{\|x^i - x^j\|_2^2}{2\mu^2}).$$

– Noyau multi quadratique inverse

$$K(x^i, x^j) = \frac{1}{\sqrt{\|x^i - x^j\|_2^2 + \beta}}.$$

Chapitre 3

Fonctions noyaux pour les SVMs

1 Introduction

Il est à signaler, que la théorie des fonctions noyaux existée avant l'apparition des SVMs. La maîtrise de cette théorie est fondamentale pour la bonne compréhension du rôle que joue cette classe de fonction dans la séparation non linéaire des données. Dans ce chapitre on va s'intéresser à quelques résultats dont on aura besoin pour la suite. La plupart des résultats seront donnés sans démonstration, le lecteur intéressé peut consulter l'ouvrage [40].

2 Fonctions noyaux

Définition 3.1 *[7] Soit* X *un ensemble non vide, une fonction k est dite fonction noyau s'il existe un* $\mathbb{R}$*-espace de Hilbert* H *et une transformation* $\Phi : X \to H$ *telle que :*

$$\begin{array}{rccc} k: & X \times X & \to & \mathbb{R} \\ & (x, x') & \longmapsto & k(x, x') = \langle \Phi(x), \Phi(x') \rangle_H \end{array}, \tag{3.1}$$

H *s'appelle espace caractéristique.*

Remarque 3.1 *[7]*

- *Dans le cas complexe, comme le produit scalaire est antisymétrique, alors l'équation (3.1) est équivalente à* $k(x, x') = \langle \phi(x), \overline{\phi(x')} \rangle$.
- *Il n'y a pas de bijection entre les fonctions noyaux et les espaces caractéristiques.*
  En effet, soient $X = \mathbb{R}$*, et* $k(x, x') = xx'$ *pour tout* $x, x' \in \mathbb{R}$*, on définit la transformation* $\Phi : x \to (\frac{x}{\sqrt{2}}, \frac{x}{\sqrt{2}})$*, alors,*

$$\langle \Phi(x), \Phi(x') \rangle_{\mathbb{R}^2} = \frac{x}{\sqrt{2}} \frac{x'}{\sqrt{2}} + \frac{x}{\sqrt{2}} \frac{x'}{\sqrt{2}} = xx' = k(x, x'),$$

mais il est facile de voir que si on définit la transformation $\Phi : x \to (\frac{x}{\sqrt{3}}, \frac{x}{\sqrt{3}}, \frac{x}{\sqrt{3}})$*, on obtient :*

$$\langle \Phi(x), \Phi(x') \rangle_{\mathbb{R}^3} = \frac{x}{\sqrt{3}} \frac{x'}{\sqrt{3}} + \frac{x}{\sqrt{3}} \frac{x'}{\sqrt{3}} + \frac{x}{\sqrt{3}} \frac{x'}{\sqrt{3}} = xx' = k(x, x'),$$

par conséquent, une fonction noyau peut être associée à plusieurs espaces caractéristiques.

Lemme 3.1 *[7] Soient* X *un ensemble non vide, et* $f_n : X \to \mathbb{R}$, $n \in \mathbb{N}$, *des fonctions telles que :* $(f_n(x)) \in \ell_2$, *pour tout* $x \in X$. *Alors,*

$$k(x,x') = \sum_{i=0}^{\infty} f_i(x) f_i(x'), \qquad x, x' \in X, \tag{3.2}$$

définit un noyau dans X.

Preuve. En utilisant l'inégalité de Hölder pour l'espace des suites ℓ_1 et ℓ_2, on obtient

$$\sum_{i=1}^{\infty} | f_n(x) f_n(x') | \leqslant \| (f_n(x)) \|_{\ell_2} \| (f_n(x')) \|_{\ell_2},$$

d'où la série (3.2) converge absolument pour $x, x' \in X$, on pose $H = \ell_2$ et on définit $\phi : X \to H$ par $\phi(x) := (f_n(x))$, $x \in X$, d'où le résultat. ■

Lemme 3.2 (Restrection des noyaux) *[7] Etant donnés un noyau* k *sur* X, $\tilde{X}$ *un ensemble, et une application* $A : \tilde{X} \to X$. *Alors,* $\tilde{k} := k(A(x), A(x'))$, $x, x' \in \tilde{X}$, *est un noyau sur* $\tilde{X}$. *En particulier, si* $\tilde{X} \subset X$, *alors* $k_{|\tilde{X}\times\tilde{X}}$ *est un noyau.*

Il est possible de construire des noyaux à partir des noyaux déjà existants, pour ce faire on donne dans la suite quelques propriétés des fonctions noyaux. On entame les propriétés par ce lemme simple.

Lemme 3.3 (Somme des noyaux) *[7] Soient* k, k_1, k_2 *des noyaux définis sur* X, *et* $\alpha \geq 0$. *Alors,* αk *et* $k_1 + k_2$ *sont aussi des noyaux définis sur* X.

Lemme 3.4 (Produit des noyaux) *[7] Soient* k_1,k_2 *deux noyaux définis sur* X_1, X_2 *réspectivement. Alors,* $k_1 \cdot k_2$ *est un noyau définis sur* $X_1 \times X_2$, *en particulier, si* $X_1 = X_2$, *alors* $k(x,x') = k_1(x,x') k_2(x,x')$, *pour tout* $x, x' \in X_1$ *est un noyau définis sur* X_1.

Avant de donner la démonstration de ce lemme, rappellons le produit tensoriel

Produit tensoriel[40]

Soient H_1 et H_2 deux espaces de Hilbert, leur somme directe $H_1 \oplus H_2$ est un espace de Hilbert, constitué des paires $(x^1, x^2) \in H_1 \times H_2$, et dont la norme définie par

$$\| (x^1, x^2) \|^2_{H_1 \oplus H_2} := \| x^1 \|^2_{H_1} + \| x^2 \|^2_{H_2}. \tag{3.3}$$

Dans le but de définir le produit tensoriel $H_1 \otimes H_2$ de H_1 et H_2, on a besoin de rappeler que pour un espace vectoriel E, une fonction $f : H_1 \times H_2 \to E$ est appelée bilinéaire, si $f(x^1, .) : H_2 \to E$ et $f(., x^2) : H_1 \to E$ sont linéaires pour tout $x^1 \in H_1$ et $x^2 \in H_2$.

Etant donné, deux espaces de Hilbert H_1 et H_2, ou d'une manière générale, deux espaces vectoriels, alors on peut montrer qu'il existe un espace vectoriel E et une fonction bilinéaire $\pi : H_1 \times H_2 \to E$ tel que pour tout espace vectoriel F et toute fonction bilinéaire $f : H_1 \times H_2 \to F$, il existe exactement une fonction linéaire $\varphi : E \to F$ tel que $f = \varphi \circ \pi$([36]). On écrit $x^1 \otimes x^2 := \pi(x^1, x^2)$ avec $(x^1, x^2) \in H_1 \times H_2$. Il se trouve que l'espace E est uniquement determiné à un isomorphisme prés, et

$$E = span\{x^1 \otimes x^2 : (x^1, x^2) \in H_1 \times H_2\}. \tag{3.4}$$

Ceci justifie le fait que $H_1 \otimes H_2 := E$ pour le produit tensoriel de H_1 et H_2.

Cependant, si H_1 et H_2 sont des espaces de Hilbert, donc il existe un seul produit scalaire $\langle .,. \rangle_{H_1 \otimes H_2}$ sur $H_1 \otimes H_2$ et sa norme correspondante $\| . \|_{H_1 \otimes H_2}$satisfait ([36]) :

$$\| x^1 \otimes x^2 \|_{H_1 \otimes H_2} = \| x^1 \|_{H_1} . \| x^2 \|_{H_2} \; ; \quad \forall (x^1, x^2) \in H_1 \times H_2.$$

En général, cet espace muni de cette norme n'est pas complet, et donc on note par $H_1 \widehat{\otimes} H_2$ la completion de $H_1 \otimes H_2$.

Preuve. (Lemme 3.4) Soient H_i l'espace caractéristique et $\phi_i : X^i \rightarrow H_i$ une fonction caractéristique de k_i, $i = 1,2$. En utilisant la définition du produit scalaire dans l'espace du produit tensoriel $H_1 \otimes H_2$ et sa completion $H_1 \widehat{\otimes} H_2$, on obtain

$$\begin{aligned} k_1(x^1,(x^1)').k_2(x^2,(x^2)') &= \langle \phi_1((x^1)'), \phi_1(x^1) \rangle_{H_1} . \langle \phi_2((x^2)'), \phi_2(x^2) \rangle_{H_2} \\ &= \langle \phi_1((x^1)') \otimes \phi_2((x^2)'), \phi_1(x^1) \otimes \phi_2(x^2) \rangle_{H_1 \widehat{\otimes} H_2} \end{aligned}$$

qui résulte que $\phi_1 \otimes \phi_2 : X_1 \times X_2 \rightarrow H_1 \widehat{\otimes} H_2$ est une fonction caratéristique de $k_1.k_2$. Pour la deuxième implication, on remarque que k est une restriction de $k_1.k_2$. ■

Théorème 3.1 *[40] soient $\mathring{B}_r(\mathbb{R})$ la boule ouverte de $\mathbb{R}$ de centre r ($r \in [0,\infty]$), et $f : \mathring{B}_r(\mathbb{R}) \longrightarrow \mathbb{R}$ une fonction développable en série de Taylor*

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \qquad x \in \mathring{B}_r(\mathbb{R}).$$

si $a_n \geq 0$ pour tout $n \in \mathbb{N}$, alors $k(x,x') = f(\langle x, x' \rangle_{\mathbb{R}^d}) = \sum_{n=0}^{\infty} a_n (\langle x, x' \rangle_{\mathbb{R}^d})^n$, $x, x' \in \mathring{B}_{\sqrt{r}}(\mathbb{R}^d)$ avec $d \in \mathbb{N}^*$ définit un noyau sur $\mathring{B}_{\sqrt{r}}(\mathbb{R}^d)$. Dans ce cas, on dit que k est un noyau de type Taylor.

À l'aide du théorème précédent,on a :

Exemple 3.1 *[40] Pour $d \in \mathbb{N}^*$, pour tout $x, x' \in \mathbb{R}^d$, on a $k(x,x') := \exp(\langle x, x' \rangle_{\mathbb{R}^d})$ est un noyau sur $\mathbb{R}^d$ appelé noyau exponentiel.*

On peut introduire maintenant un des noyaux les plus utilisés en pratique, il est donné par la proposition suivante :

Proposition 3.1 (Noyau RBF) *[39] Pour $d \in \mathbb{N}$, $\gamma > 0$, $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$ et $x' = (x_1', \cdots, x_d') \in \mathbb{R}^d$, on définie :*

$$k_\gamma(x,x') = \exp(-\gamma^{-2} \sum_{i=1}^{d} (x_i - x_i')^2).$$

Alors, k_γ est un noyau sur $\mathbb{R}^d$ appelé fonction à base radial (RBF) ou bien noyau Gaussien, il peut s'écrire aussi :

$$k_\gamma(x,x') = \exp(-\frac{\|x - x'\|_2^2}{\gamma^2}), \qquad x, x' \in \mathbb{R}^d.$$

Preuve. Etant donné $x, x' \in \mathbb{R}^d$ fixés, en décomposant $k_{\gamma,\mathbb{R}^d}$ on aura :

$$k_{\gamma,\mathbb{R}^d}(x,x') = \frac{\exp(2\gamma^{-2}\langle x,x'\rangle)}{\exp(\gamma^{-2}\sum_{j=1}^{d} x_j^2)\exp(\gamma^{-2}\sum_{j=1}^{d} x_j'^2)}$$

et en appliquant les lemmes (3.4 et 3.2) et l'exemple (3.1), on trouve la première implication. La deuxième est triviale. ■

On donne à présent une caractérisation pratique des noyaux. D'abord, on aura besoin de la définition suivante :

Définition 3.2 *[40] Une fonction $k : X \times X \longrightarrow \mathbb{R}$ est dite semi-définie positive si, pour tout $n \in \mathbb{N}$, pour tous $\alpha_1, \cdots, \alpha_n \in \mathbb{R}$ et pour tous $x_1, \ldots, x_n \in X$, on a*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0,$$

k est dite définie positive si, pour tout $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n$, tel que $\alpha \neq 0$ et $x_1, \ldots, x_n \in X$, mutuellement distincts, on a

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) > 0,$$

finalement, k est dite symétrique si, $k(x,x') = k(x',x)$, pour tout $x, x' \in X$.

Si on définit la matrice de Gram K d'un noyau k par rapport aux éléments $x_1, \ldots, x_n \in X$ comme $\left[K_{ij}\right]_{i,j=1,\cdots,n} = \left[k(x_i,x_j)\right]_{i,j=1,\ldots,n}$, alors la définition précédente est équivalente à la semi-définie positivité (au sens matriciel) de K pour tout $x_1, \ldots, x_n \in X$.

Comme k est a valeurs réelles, avec la fonction caractéristique $\phi : X \to H$, alors k est symétrique (car le produit scalaire est symétrique). En plus, k est aussi semi-définie positive car pour tout $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ et $x_1, \ldots, x_n \in X$, on a :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \sum_{j=1}^{n} \alpha_j \phi(x_j) \rangle_H \geqslant 0.$$

3 Théorème de Mercer

Un dernier résultat qui permet de caractériser les fonctions noyaux sans passer explicitement par l'espace caractéristique est donné par le théorème suivant :

Théorème 3.2 (Théorème de Mercer) *[7] Une fonction $k : X \times X \longrightarrow \mathbb{R}$ est un noyau si est seulement si k est symétrique et semi-définie positive.*

Preuve. Pour montrer qu'une fontion k symétrique et semi-définie positive est un noyau, on pose

$$\begin{aligned} H_{pre} &= \{\sum_{i=1}^{n} \alpha_i k(.,x_i) : n \in \mathbb{N},\ \alpha_1, \ldots, \alpha_n \in \mathbb{R},\ x_1, , \ldots, x_n \in X\}, \\ f &:= \sum_{i=1}^{n} \alpha_i k(.,x_i) \in H_{pre} \quad \text{et} \\ g &:= \sum_{j=1}^{m} \beta_j k(.,x_i') \in H_{pre}. \end{aligned}$$

On définit aussi

$$\langle f, g\rangle := \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j k(x_i', x_j). \tag{3.5}$$

Notons que (3.5) est indépendante de g car $\langle f, g\rangle = \sum_{j=1}^{m}\beta_j f(x_j')$. En outre, comme k est symétrique, on trouve $\langle f, g\rangle = \sum_{i=1}^{n}\alpha_i g(x_i)$ c-à-d (3.5) est aussi indépendante de f. De plus, (3.5) montre que $\langle .,.\rangle$ est bilinéaire et symétrique, et comme k est semi-définie positive, $\langle .,.\rangle$ est aussi semi-définie positive, i.e. $\langle f, f\rangle \geqslant 0$ pour tout $f \in \mathrm{H}_{pre}$. Par conséquent, $\langle .,.\rangle$ vérifie l'inégalité de Cauchy-Schwarz i.e.,

$$|\langle f, g\rangle|^2 \leqslant \langle f, f\rangle.\langle g, g\rangle, \qquad f, g \in \mathrm{H}_{pre}.$$

Posons $f := \sum_{i=1}^{n}\alpha_i k(., x_i) \in \mathrm{H}_{pre}$ avec $\langle f, f\rangle = 0$, puis pour tout $x \in \mathrm{X}$, on a

$$|f(x)|^2 = |\sum_{i=1}^{n}\alpha_i k(x, x_i)|^2 = |\langle f, k(., x)\rangle|^2 \leqslant \langle k(., x), k(., x)\rangle.\langle f, f\rangle = 0,$$

et donc on trouve $f = 0$. Donc, on a montré que $\langle .,.\rangle$ est un produit scalaire dans H_{pre}. Soient H la completion de H_{pre} et $\mathrm{I} : \mathrm{H}_{pre} \to \mathrm{H}$ l'injection canonique. Alors H est un espace de Hilbert et on a

$$\langle \mathrm{I}k(., x'), \mathrm{I}k(., x)\rangle_{\mathrm{H}} = \langle k(., x'), k(., x)\rangle_{\mathrm{H}_{pre}} = k(x, x'),$$

pour tout $x, x' \in \mathrm{X}$, i.e., $x \longmapsto \mathrm{I}k(., x)$, $x \in \mathrm{X}$, définit une fonction de transformation du noyau k ■

4 Espace de Hilbert à noyau reproduisant (RKHS ou « Reproducing Kernel Hilbert Space »).

Dans cette partie, nous introduisons les espaces de Hilbert à noyau reproduisant (RKHS) en decrivant leur relation avec les fonctions noyaux. En particuler, on verra que dans un certain sens le RKHS d'une fonction noyau est le plus petit espace caractéristique de ce noyau. Par conséquent, on peut le considérer comme l'espace caractéristique canonique.

Définition 3.3 (Espace de Hilbert à Noyau Reproduisant) *[40] Soient* $\mathrm{X} \neq 0$ *et* H *un espace de Hilbert de fonctions réelles définies sur l'ensemble* X *à valeurs réelles.*

- *Une fonction* $k : \mathrm{X} \times \mathrm{X} \to \mathbb{R}$ *est un noyau reproduisant de* H, *si on a :* $k(., x) \in \mathrm{H}$, *pour tout* $x \in \mathrm{X}$ *et la propriété :*

$$f(x) = \langle f, k(., x)\rangle$$

  est vrai pour tout $f \in \mathrm{H}$ *et* $x \in \mathrm{X}$.
- *L'espace* H *est appelé "espace de Hilbert à noyau reproduisant (RKHS)" sur* X, *si pour tout* $x \in \mathrm{X}$, *la fonction de Dirac* $\delta_x : \mathrm{H} \to \mathbb{R}$ *définie par :*

$$\delta_x(f) = f(x), \qquad f \in \mathrm{H}$$

  est continue.

Lemme 3.5 (Les noyaux reproduisant sont des noyaux) *[7] Soient k un noyau reproduisant et* H *son RKHS. La fonction caractéristique* $\phi : X \to H$ *donnée par*

$$\phi(x) = k(.,x), \qquad x \in X$$

est appelée fonction caractéristique canonique.

Théorème 3.3 (Chaque RKHS à un unique noyau reproduisant) *[7] Soit* H *un RKHS sur* X. *Alors* $k : X \times X \to \mathbb{R}$ *définie par :*

$$k(x,x') = \langle \delta_x, \delta_{x'} \rangle_H, \qquad x, x' \in H$$

est l'unique noyau reproduisant de H. *de plus, si* $(e_i)_{i\in I}$ *est une base orthonormale de* H, *alors pour tout* $x, x' \in H$, *on a :*

$$k(x,x') = \sum_{i\in I} e_i(x) e_i(x').$$

Théorème 3.4 (Chaque noyau à un unique RKHS) *[7] Soient* X *un ensemble non vide, k une fonction noyau sur* X *avec espace caractéristique* H_0 *et une fonction caractéristique* $\phi_0 : x \to H_0$. *Alors*

$$H := \{ f : X \to \mathbb{R} \mid \exists w \in H_0 \text{ avec } f = \langle w, \phi_0(.) \rangle_{H_0} \ , \ \forall x \in X \} \tag{3.6}$$

équipé de la norme

$$\| f \|_H := \inf\{ \| w \|_{H_0} : w \in H_0 \text{ avec } f = \langle w, \phi_0(.) \rangle_{H_0} \} \tag{3.7}$$

est le seul RKHS pour lequel k est un noyau reproduisant.

Par conséquent, les équations (3.6) et (3.7) sont indépendantes du choix de H_0et ϕ_0. Comme, l'opérateur $V : H_0 \to H$ définit par

$$Vw := \langle w, \phi_0(.) \rangle_{H_0}, \qquad w \in H_0$$

est une distance surjective, i.e., $V\overset{\circ}{B}_{H_0} = \overset{\circ}{B}_H$ tel que $\overset{\circ}{B}_{H_0}$ et $\overset{\circ}{B}_H$sont les boules d'unités ouvertes de H_0 et H respectivement. Finalement, l'ensemble

$$H_{pre} := \{ \sum_{i=1}^{n} \alpha_i k(.,x_i) : n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_n \in X \}$$

est dense dans H, et pour $f := \sum_{i=1}^{n} \alpha_i k(.,x_i) \in H_{pre}$, on a

$$\| f \|_H^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j).$$

Notre but maintenant est de déterminer une formule explicite du RKHS du noyau Gaussien (RBF). Pour ce faire, on définit $\gamma > 0$ et $d \in \mathbb{N}^*$, pour $f : \mathbb{C}^d \to \mathbb{C}$ une fonction holomorphe donnée, on définit

$$\| f \|_{\gamma,\mathbb{C}^d} := \left(\frac{2^d}{\pi^d \gamma^{2d}} \int_{\mathbb{C}^d} | f(z) |^2 \, e^{\gamma^{-2} \sum_{j=1}^{d} (z_j - \overline{z}_j)^2} \, dz \right)^{\frac{1}{2}},$$

ou z_j est la j ème composante de $z \in \mathbb{C}^d$ et $\overline{z}_j$ son conjugué. alors, on écrit :

$$H_{\gamma,\mathbb{C}^d} := \{ f : \mathbb{C}^d \to \mathbb{C} \mid f \ holomorphe \ et \ \| f \|_{\gamma,\mathbb{C}^d} < \infty \}.$$

Donc $H_{\gamma,\mathbb{C}^d}$ est un espace vectoriel dans $\mathbb{C}$ muni de la norme $\| \, . \, \|_{\gamma,\mathbb{C}^d}$ est un espace pre-Hilbertien.

Théorème 3.5 (RKHS du Gaussian (RBF) complexe) *[40] Soient* $\gamma > 0$ *et* $d \in \mathbb{N}$. *Alors* $(\mathrm{H}_{\gamma,\mathbb{C}^d}, \| \cdot \|_{\mathrm{H}_{\gamma,\mathbb{C}^d}})$ *est un RKHS et* $k_{\gamma,\mathbb{C}^d}$ *son noyau reproduisant. De plus, pour* $n \in \mathbb{N}_0$, *on a* $e_n : \mathbb{C} \to \mathbb{C}$ *est défini par :*

$$e_n(z) := \sqrt{\frac{2^n}{\gamma^{2n} n!}} z^n e^{-\gamma^{-2} z^2}, \quad z \in \mathbb{C}.$$

Donc le système $(e_{n_1} \otimes \cdots \otimes e_{n_d})_{n_1,\ldots,n_d \geqslant 0}$ *des fonctions* $e_{n_1} \otimes \cdots \otimes e_{n_d} : \mathbb{C}^d \to \mathbb{C}$, *défini par*

$$e_{n_1} \otimes \cdots \otimes e_{n_d}(z_1, \ldots, z_d) := \prod_{j=1}^{d} e_{n_j}(z_j), \qquad (z_1, \cdots, z_d) \in \mathbb{C}^d,$$

est une base orthonormale de $\mathrm{H}_{\gamma,\mathbb{C}^d}$.

5 Noyaux universels

Définition 3.4 *[39] Un noyau* k *continu sur un espace métrique* X *est appellé universel si son RKHS* H *est dense dans* C(X), *i.e., pour toute fonction* $g \in \mathrm{C(X)}$, *et* $\forall \varepsilon > 0$, *il existe une fonction* $f \in \mathrm{H}$, *telle que*

$$\| f - g \|_\infty \leqslant \varepsilon.$$

Dans la suite nous allons discuter quelques propriétées géométriques des noyaux universels. Pour ce faire, nous aurons besoin des définitions suivantes :

Définition 3.5 *[39] Soit* k *un noyau sur un espace métrique* X *avec RKHS* H.*On dit que* k *sépare deux ensembles disjoints* $\mathrm{A}, \mathrm{B} \subset \mathrm{X}$, *s'il existe une fonction* $f \in \mathrm{H}$, *avec* $f(x) > 0$ *pour tout* $x \in \mathrm{A}$ *et* $f(x) < 0$ *pour tout* $x \in \mathrm{B}$. *Aussi on dira que* k *sépare tout ensemble fini (ou compact) si* k *sépare tout ensemble fini disjoint (ou compact)* $\mathrm{A}, \mathrm{B} \subset \mathrm{X}$.

Théorème 3.6 *[40] Soit* X *un ensemble compact métrique et* k *un noyau universel sur* X. *Alors* k *sépare tout les ensembles compacts disjoints.*

Preuve. Soient A, B deux sous ensembles compacts disjoints de X et d une distance sur X. pour tout $x \in \mathrm{X}$, on définit :

$$g(x) := \frac{dist(x,\mathrm{B})}{dist(x,\mathrm{A}) + dist(x,\mathrm{B})} - \frac{dist(x,\mathrm{A})}{dis(x,\mathrm{A}) + dist(x,\mathrm{B})},$$

on utilise la fonction de distance $dist := \inf_{x' \in \mathrm{C}} dist(x, x')$, pour $x \in \mathrm{X}$ et $\mathrm{C} \subset \mathrm{X}$. Comme A et B sont compacts et la fonction distance est continue, g l'est aussi. De plus, $g(x) = 1$, pour tout $x \in \mathrm{A}$ et $g(x) = -1$, pour tout $x \in \mathrm{B}$. Soit H le RKHS de la fonction noyau k. Alors, il existe une fonction $f \in \mathrm{H}$ telle que $\| f - g \|_\infty \leqslant \frac{1}{2}$; et donc par construction de la fonction g, on a : $f(x) \geqslant \frac{1}{2}$ pour tout $x \in \mathrm{A}$ et $f(x) \leqslant \frac{1}{2}$ pour tout $x \in \mathrm{B}$. ■

Bien que la proposition (3.6) découle de la notion d'universalité, elle a des conséquences surprenantes pour l'interprétation géométrique des fonctions caractéristiques de la fonction noyau universel. En effet, soit k un noyau universel sur X avec espace caractéristique H_0 et fonction caractéristique $\phi_0 : \mathrm{X} \to \mathrm{H}_0$. En outre, supposons qu'il existe un sous ensemble fini $\{x^1, \cdots, x^n\} \subset \mathrm{X}$. Alors le théorème (3.6) assure que pour n'importe

quel choix des signes $y_1, \cdots, y_n \in \{-1, 1\}$, on trouve une fonction f dans l'espace RKHS H avec $y_i f(x^i) > 0$, $i = 1, \cdots, n$. De l'équation (3.4), la fonction f peut être représentée par $f = \langle w, \phi_0(.)\rangle$ pour un approprié $w \in H_0$. Par conséquent, l'ensemble des données transformées $((\phi_0(x^1), y_1), \cdots, (\phi_0(x^n), y_n))$ peut être correctement séparé en H_0 par l'hyperplan défini par w.

Corollaire 3.1 (Exemples de noyaux universels) *[39] Soient* X *un ensemble compact de* $\mathbb{R}^d$, $\gamma > 0$ *et* $\alpha > 0$. *Alors les noyaux donnés ci-dessus sont des noyaux universels sur* X.

$$\begin{aligned} \textit{noyau exponentiel} \quad &: \quad k(x, x') := \exp(\langle x, x'\rangle), \\ \textit{noyau Gaussien RBF} \quad &: \quad k_\gamma(x, x') := \exp(-\gamma^{-2} \parallel x - x' \parallel_2^2, \\ \textit{noyau polynomial} \quad &: \quad k(x, x') := (1 - \langle x, x'\rangle)^{-\alpha}, \end{aligned}$$

De plus, pour le dernier noyau nous supposons que $X \subset \{x \in \mathbb{R}^d : \parallel x \parallel_2 < 1\}$.

Chapitre 4

Analyse du Noyau de Legendre pour les SVMs

1 Polynômes de Legendre

Définition 4.1 *[16] Les polynômes de Legendre, noté* P_n ; $n \geqslant 0$, *sont des polynômes de dergé n, définis sur l'intervalle* $[-1,1]$, *comme étant les solutions de l'équation différentielle, dite de Legendre :*

$$(1-x^2)\frac{d^2y}{dx^2}(x) - 2x\frac{dy}{dx}(x) + n(n+1)y(x) = 0; \qquad -1 < x < 1.$$

Proposition 4.1 (Caractérisation et propriétés) *[16] La famille des polynômes de Legendre* $\{P_n\}_{n\geqslant 0}$ *est caractérisée par l'une des trois relations suivantes :*

1. *Ils sont définis par la formule dite de Rodriguez :*

$$P_n(x) = \frac{1}{n!2^n}\frac{d^n}{dx^n}(x^2-1)^n \; ; \; n \geqslant 0.$$

2. *Ils vérifient les relations de récurence :*

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \; ; \; \forall n \geqslant 1,$$

   avec

$$P_0(x) = 1 \; et \; P_1(x) = x.$$

3. *Ils admettent la représentation intégrale :*

$$P_n(x) = \frac{1}{\pi}\int_0^{\pi}[x + \sqrt{x^2-1}\cos(t)]^n dt.$$

   Ces polynômes possèdent la proprétés suivantes :

Proposition 4.2 *[16] La famille des polynômes de Legendre* $\{P_n\}_{n\geqslant 0}$ *est orthogonale dans* $L^2([-1,1])$:

$$\langle P_n, P_m\rangle = \int_{-1}^{1} P_n(x)P_m(x)dx = \begin{cases} 0 & si \; n \neq m \\ \frac{2}{2n+1} & si \; n = m \end{cases} \tag{4.1}$$

Le polynôme de Legendre P_n *est pair (respectivement impair) si n est pair (respectivement impair), de plus :*

$$P_n(1) = 1 \; et \; P_n(-1) = (-1)^n, \; \forall n \geqslant 0.$$

Théorème 4.1 *[16] Soit h une fonction de carré intégrable sur* $[-1,1]$*, alors, on a dans* $L^2(]-1,1[)$

$$h = \sum_{n\geqslant 0} \frac{2n+1}{2} \langle h, P_n \rangle P_n, \tag{4.2}$$

et on a l'égalité de $Bassel - Parseval$

$$\| h \|^2 = \int_{-1}^{1} | h(x) |^2 \, dx = \sum_{n\geqslant 0} \frac{2n+1}{2} | \langle h, P_n \rangle |^2 .$$

Remarque 4.1 *Si on pose* $L_n(x) = \sqrt{\frac{2n+1}{2}} P_n(x)$*, alors la famille* $\{L_n\}_{n\geqslant 0}$ *est orthonormé dans* $L^2([-1,1])$.

Dans la section suivate, nous allons voir qu'on peut exprimer une fonction f sous la forme (4.2), sans que la famille $\{L_n\}_{n\geqslant 0}$ soit orthonormale.

2 Les Frames

2.1 Introduction

La théorie des Frames a été introduite par Schaeffer (1952) (Daubechies, 1992) dans le but d'établir des conditions générales sous lesquelles on peut construire parfaitement une fonction f dans un espace de Hilbert H à partir de son produit scalaire $(\langle .,. \rangle_H)$ avec une famille de vecteurs $\{\varphi_n\}_{n\in\Gamma}$ où Γ est un ensemble d'indixes fini ou infini dénombrable. Dans la section suivante, nous allons voir qu'on peut exprimer une fonction f sous la forme (4.2), sans que la famille $(L_n)_n$ soit orthonormale. Nous verrons aussi un résultat important qui caractérise à partir d'une famille de Frame donnée, l'espace reproduisant associé a ce noyau.

Définition 4.2 *[35] Une famille de vecteurs* $\{\varphi_n\}_{n\in\Gamma}$ *est un frame d'espace de Hilbert* H*, s'il existe deux constantes* $A > 0$ *et* $0 < A \leqslant B < \infty$ *tel que*

$$\forall f \in H, \; A \| f \|_H^2 \leqslant \sum_{n\in\Gamma} | \langle f, \varphi_n \rangle_H |^2 \leqslant B \| f \|_H^2 .$$

Le frame est dit serré si A *est égal à* B.

Si l'ensemble $\{\varphi_n\}_{n\in\Gamma}$ satisfait les conditions des frames, alors l'opérateur du frame U peut étre défini par

$$U : H \to \ell^2$$
$$f \to \{\langle f, \varphi_n \rangle_H\}_{n\in\Gamma}$$

La reconstruction de f à partir de ses coéfficients de frame nécessite la définition d'un frame dual. A cet effet, on introduit l'opérateur adjoint U^* de U qui existe et est unique parce qu'il se trouve sur un espace de Hilbert.

$$U^* : \ell^2 \to H$$
$$\{c_n\}_{n\in\Gamma} \to \sum_{n\in\Gamma} c_n \varphi_n .$$

Théorème 4.2 *[35] [Daubechies 1992] Soit* $\{\varphi_n\}_{n\in\Gamma}$ *un frame de* H *avec les bornes de frames* A *et* B. *On définit le frame dual* $\{\overline{\varphi}_n\}_{n\in\Gamma}$ *comme* $\overline{\varphi}_n = (U^*U)^{-1}\varphi_n$.
Pour tout $f \in H$ *on a :*

$$\frac{1}{B}\parallel f\parallel_H^2 \leqslant \sum_{n\in\Gamma} |\langle f, \overline{\varphi_n}\rangle_H|^2 \leqslant \frac{1}{A}\parallel f\parallel_H^2,$$

et

$$f = \sum_{n\in\Gamma}\langle f, \overline{\varphi}_n\rangle_H \varphi_n = \sum_{n\in\Gamma}\langle f, \varphi_n\rangle_H \overline{\varphi}_n$$

Si le frame est serré, alors $\overline{\varphi}_n = \frac{1}{A}\varphi_n$.

On verra aussi un résultat important qui caractérise l'espace reproduisant associé au noyau à partir d'une famille de frame donné.

Théorème 4.3 *[35] Soient* H *un espace de Hilbert à noyau reproduisant et* $H \in Hilb(\mathbb{R}^{\Omega})$ *et la famille* $\{\varphi_n\}_{n\in\Gamma}$ *est un frame de cet espace. Alors, le noyau reproduisant* $k(s,t)$ *est définit par :*

$$\begin{aligned} k &: \Omega\times\Omega \to \mathbb{R} \\ s\times t &\to k(s,t) = \sum_{n\in\Omega}\overline{\varphi}_n(s)\varphi_n(t). \end{aligned}$$

Théorème 4.4 *[35] Soient* $N \in \mathbb{N}$ *et* $\{\varphi_n\}_{n=1,\cdots,N}$ *un ensemble fini de fonctions non nulles d'un espace de Hilbert* $(B, \langle .,.\rangle)$ *avec* $B \in \mathbb{R}^{\Omega}$ *tel que*

$$\exists M,\ \forall t\in\Omega,\ \forall n\ \ 1\leqslant n\leqslant N,\quad |\varphi_n(t)| < M.$$

Soit H *l'ensemble de fonctions tel que*

$$H = \{f = \sum_{n=1}^{N} a_n\varphi_n :\ a_n\in\mathbb{R},\ n = 1,\ldots,N\}.$$

Alors, $(H, \langle .,.\rangle_B)$ *est un RKHS et son noyau reproduisant est*

$$k(s,t) = \sum_{n=1}^{N}\overline{\varphi}_n(s)\varphi_n(t),$$

avec $\{\overline{\varphi}_n\}_{n=1,\ldots,N}$ *est le frame dual de* $\{\varphi_n\}_{n=1,\ldots,N}$ *de* H.

3 Noyaux de Legendre

Définition 4.3 (Fonction noyau de Legendre) *[32] Soit* $k : [-1,1]\times[-1,1]\to\mathbb{R}$ *défini par*

$$k(s,t) = \sum_{n=0}^{N} L_n(s)\,L_n(t). \tag{4.3}$$

Pour tout $t_1,\cdots,t_m \in [-1,1]$, la matrice de Gram $K = (k(t_i,t_j))_{i,j}$, $i,j = 1,\cdots,m$, est semi-definie positive. D'aprés le théorème de Mercer (3.2), la fonction (4.3) est une fonction noyau, appelée noyau de Legendre (le paramétre N est le plus grand degrés du polynome de Legendre). La fonction de transformation associée a ce noyau est

$$t \mapsto \phi_N(t) = (L_0(t), L_1(t), \cdots, L_N(t)) \in \mathbb{R}^{N+1}. \tag{4.4}$$

Il est clair que pour $N \in \mathbb{N}$ fixé, il existe une constante positive M tel que

$$\forall t \in [-1,1],\ \forall n\ 0 \leqslant n \leqslant N,\ |\ L_n(t)\ | < M. \tag{4.5}$$

Cette propriété va nous permettre d'affirmer le théorème suivant dont la preuve a été omis puisque c'est une conséquence directe du théorème (4.4)

Théorème 4.5 *[11] Soient* $N \in \mathbb{N}^*$, $\{L_n\}_{n\in\mathbb{N}}$ *définie par (4.3) et* H *un ensemble de fonctions tel que*

$$H = \{f = \sum_{n=1}^{N} a_n \phi_n :\ a_n \in \mathbb{R},\ n = 1, \cdots, N\}. \tag{4.6}$$

Alors, $(H, \langle .,. \rangle_{L^2[-1,1]})$ *est un RKHS et son noyau reproduisant est (4.3).*

Pour donner une généralisation du noyau (4.3), dans le cas multidimensionnel ($\mathbf{x}^i \in \mathbb{R}^d$ pour $i = 1, \cdots, m$) et comme les polynomes de Legendre sont définis sur $[-1,1]$ une procedure de mise a l'échelle est nécessaire. Par exemple la transformation suivante :

$$x_j^i := \frac{2\left(x_j^i - \min_j(x_j^i)\right)}{\max_j\left(x_j^i\right) - \min_j(x_j^i)} - 1 \quad \text{pour } j = 1, \cdots, d. \tag{4.7}$$

transforme chaque donnée $i = 1, \ldots, m$ dans l'intervalle $[-1,1]$.

Pour ce qui suit, on considère l'ensemble $X = [-1,1]^d$. On définit une fonction $\mathcal{K} : X \times X \to \mathbb{R}$ comme suit

$$\mathcal{K}\left(\mathbf{x},\mathbf{y}\right) = \prod_{i=1}^{d} k(x_i, y_i) = \prod_{i=1}^{d} \sum_{n=0}^{N} L_n(x_i)\, L_n\left(y_i\right). \tag{4.8}$$

On introduit l'ensemble

$$\mathcal{H} = H^{\otimes d} = \underbrace{H \otimes H \otimes \cdots \otimes H}_{d\ fois}, \tag{4.9}$$

appelé produit tensoriel d'orde d de H([36]). Il est clair que $\mathcal{H}$ est un espace de Hilbert à valeurs réelles. Pour $f, g \in \mathcal{H}$ tel que $f = \otimes_{i=1}^{d} f_i = f_1 \otimes f_2 \otimes \cdots \otimes f_d$ et $g = \otimes_{i=1}^{d} g_i$, avec $f_i, g_i \in H$, pour $i = 1, \ldots, d$, le produit scalaire dans $\mathcal{H}$ est défini comme suit

$$\langle f, g \rangle_{\mathcal{H}} = \left\langle \otimes_{i=1}^{d} f_i, \otimes_{i=1}^{d} g_i \right\rangle_{\mathcal{H}} = \Pi_{i=1}^{d} \langle f_i, g_i \rangle_{L_2([-1,1]}. \tag{4.10}$$

Pour $\left(\mathbf{x},\mathbf{y}\right) \in X$, on a

$$\mathcal{K}\left(\mathbf{x},\mathbf{y}\right) = \Pi_{i=1}^{d} k(x_i, y_i) = \Pi_{i=1}^{d} \left\langle \phi(x_i), \phi\left(y_i\right)\right\rangle_{L_2([-1,1]} = \left\langle \otimes_{i=1}^{d} \phi(x_i), \otimes_{i=1}^{d} \phi\left(y_i\right)\right\rangle_{\mathcal{H}}. \tag{4.11}$$

On définit la fonction $\Phi : X \to \mathcal{H}$ comme suit

$$\Phi(\mathbf{x}) = \otimes_{i=1}^{d} \phi(x_i), \tag{4.12}$$

en substituant (4.12) dans (4.11) on aura

$$\mathcal{K}\left(\mathbf{x},\mathbf{y}\right) = \left\langle \Phi(\mathbf{x}), \Phi\left(\mathbf{y}\right)\right\rangle_{\mathcal{H}},$$

donc $\mathcal{K}$ est un noyau par définition, sa transormation est Φ et son espace caractéristique est $\mathcal{H}$.

La proposition suivante montre que $\mathcal{K}$ est un noyau et $\mathcal{H}$ est son RKHS.

Proposition 4.3 *[11] L'espace* H *de Hilbert sur le corps* $\mathbb{R}$*, défini par (4.9) est un RKHS dont son produit scalaire est defini par (4.10).*

Preuve.

– Soient $\mathbf{x},\mathbf{y}\in \mathrm{X}$, de la définition du produit tensoriel ([36]), on a

$$\otimes_{i=1}^{d} k(\cdot,x_i)\left(y_i\right)=\Pi_{i=1}^{d} k(x_i,y_i)=\mathcal{K}\left(\mathbf{x},\mathbf{y}\right),$$

et comme H est un RKHS, on déduit que

$$\mathcal{K}(\cdot,\mathbf{x})=\otimes_{i=1}^{d} k(\cdot,x_i)\in\mathcal{H}.$$

– Pour démontrer la propriétées de reproduction, soit $f\in\mathcal{H}$, puis $\forall\mathbf{x}\in \mathrm{X}$

$$f(\mathbf{x})=(\otimes_{i=1}^{d} f_i)(\mathbf{x})=\Pi_{i=1}^{d} f_i(x_i)=\Pi_{i=1}^{d}\left\langle f_i,k(\cdot,x_i)\right\rangle_{\mathrm{L}_2([-1,1]},$$

la première égalité est donnée par la définition du produit tensoriel des fonctions et la seconde du fait que H est RKHS. En utilisant (4.10) on déduit que

$$f(\mathbf{x})=\left\langle \otimes_{i=1}^{d} f_i,\otimes_{i=1}^{d} k(\cdot,x_i)\right\rangle_{\mathcal{H}}=\left\langle f,\mathcal{K}(\cdot,\mathbf{x})\right\rangle_{\mathcal{H}}.$$

– Comme la dimension de $\mathcal{H}$ est finie, la function de Dirac $\delta_{\mathbf{x}}$ est continue comme une fonction linéaire.

■

Remarque 4.2 *Il est clair que* H *défini par relation (4.6) n'est rien d'autre que l'espace vectoriel des polynômes de degré inférieur ou égal à* N *à une indéterminée dans* $[-1,1]$ *et* $\dim(\mathrm{H})=\mathrm{N}+1$*. Cependant,* H *est un RKHS par rapport au produit scalaire de* $\mathrm{L}^2[-1,1]$*, Où* $\{\mathrm{L}_n\}_{n=1,\cdots,\mathrm{N}}$ *forment une base orthonormale. Définissons l'ensemble* $\Im=\{\mathrm{I}_1,\cdots,\mathrm{I}_{(\mathrm{N}+1)^d}\}$ *de* $d-uplets$ *qui peuvent être construits à partir de l'ensemble* $\{0,\cdots,\mathrm{N}\}$*. Selon la théorie du produit tensoriel ([36]), l'ensemble des vecteurs* $\{\otimes_{i\in \mathrm{I}_j}\mathrm{L}_i\}_{j=1,\cdots,(\mathrm{N}+1)^d}$ *forment une base de* H*. Le produit scalaire (4.10) permet à cette base d'hériter de la propriété d'orthonormalité de* H*. Cette orthonormalité forte montrée par les tests numériques, permet au noyau défini par (4.8) de réduire la redondance de données originales et le rendre capable d'extraire des caractéristiques plus discriminantes. Une autre propriété fondamentale pour la classification non linéaire du noyau est donnée dans le résultat suivant.*

Théorème 4.6 *[11] Soient* $\mathrm{X}=[-1,1]^d$*,* $\mathcal{K}$ *et* $\mathcal{H}$ *définis par (4.8) et (4.9) respectivement. Soient* A *et* B *deux sous ensembles fermés de* X *tel que* $\mathrm{A}\cap\mathrm{B}=\emptyset$*. Alors il existe une fonction* $p\in\mathcal{H}$ *avec* $p(\boldsymbol{x})>0$ *pour tout* $\boldsymbol{x}\in\mathrm{A}$ *et* $p(\boldsymbol{x})<0$ *pour tout* $\boldsymbol{x}\in\mathrm{B}$*.*

Preuve. Soient $\mathrm{A},\mathrm{B}\subset\mathrm{X}$ deux sous ensembles fermés et d une fonction de distance telle que

$$d(\mathbf{x},\mathrm{S})=\inf_{\mathbf{s}\in\mathrm{S}}\|\mathbf{x}-\mathbf{s}\|_{\mathbb{R}^d},\ \text{pour } \mathbf{x}\in\mathrm{X}\text{ et }\mathrm{S}\subset\mathrm{X}.$$

On définit une fonction g sur X comme suit

$$g(\mathbf{x})=\frac{d(\mathbf{x},\mathrm{B})}{d(\mathbf{x},\mathrm{A})+d(\mathbf{x},\mathrm{B})}-\frac{d(\mathbf{x},\mathrm{A})}{d(\mathbf{x},\mathrm{A})+d(\mathbf{x},\mathrm{B})}.$$

On sait que $g \in C(X)$ car A et B sont compact et la fonction de distance d est continue. De plus, on a

$$g(\mathbf{x}) = \begin{cases} 1 \text{ if } \mathbf{x} \in A \\ -1 \text{ if } \mathbf{x} \in B \end{cases}.$$

En utilisant le théorème d'approximation de Weierstrass, on peut affirmer que pour tout $\varepsilon > 0$ il est existe un polynome $p \in \mathbb{R}[\mathbf{x}]$ de degrés $l \in \mathbb{N}$ tel que $\|g - p\|_\infty < \varepsilon$. Soit N le plus petit entier satisfaisant $l \leq (N+1)^d$. D'aprés la remarque (4.2) , p appartient au RKHS de $\mathscr{H}$. Pour $\varepsilon = \frac{1}{2}$, on deduit que pour tout $\mathbf{x} \in A$, $p(\mathbf{x}) > 0$ et $p(\mathbf{x}) < 0$ pour tout $\mathbf{x} \in B$. ■

Remarque 4.3 *La marge souple des SVMs vise à séparer un sous ensemble fini* $x^1, \ldots, x^m$ *de* X. *Le théorème (4.6) nous dit que pour tout choix de signes* $y_1, \ldots, y_m \in \{-1, 1\}$, *on peut trouver une fonction* p *dans le RKHS* H *du noyau* k *avec* $y_i p(x^i) >$ *pour tout* $i = 1, \cdots, m$, *ça indique que l'universalité est une condition suffisante ([39]) mais pas nécessaire pour garantir l'existence d'un hyperplan de séparation dans l'espace des caractéristiques.*

4 Expériences Numériques

Afin de montrer les puissantes propriétés de séparation du noyau défini par (4.8), nos tests numériques ont été réalisés sur deux types d'exemples. Dans le premier type d'exemples, nous considérons les ensembles de données dans un espace à deux dimensions afin de donner une comparaison géométrique entre le noyau de Legendre, RBF et polynomial. Le second type d'exemples sont cinq problèmes test sélectionnés dans le référentiel d'apprentissage automatique UCI ([21]), où les trois noyaux mentionnés ci-dessus sont comparés en ce qui concerne leur exactitude, le temps d'exécution et le nombre de vecteurs de support. La boîte à outils SVM que nous avons utilisée pour le premier type est le solveur "$fitcsvm$" de Matlab ([26]) et pour la seconde, libsvm ([6]) qui est disponible en tant que "MEX function" sous Matlab. Le programme a été implémenté moyennant Matlab R2014 sur un pc Windows 8 avec mémoire de 6G.

4.1 Comparison Géométrique

Nous considérons les exemples académiques bidimensionnels souvent utilisés pour la classification binaire non linéaire ; les ensembles de données sont twospirales et checker. L'ensemble de données twospirale a 200 données, 100 pour chaque classe (Figure 4.1). L'ensemble de données du checker comprend 1000 points de $\mathbb{R}^2$; des points rouges et bleus distribués sur les carrés d'un damier (figure 4.2).

Pour les deux exemples, les données sont d'abord normalisées par la relation (4.7), puis un ensemble d'apprentissage aléatoire de la taille de 90 % des données a été choisi et séparé des 10 % restants que nous avons considéré comme ensemble tests. Plusieurs SVMs ont été formés sur les données d'apprentissage en utilisant la validation croisée 10 fois. Le paramètre de compromis C et la valeur du paramètre RBF μ sont égal à 2^i avec $i = -15, \cdots, 0, \cdots 15$ et $i = -5, \cdots, 0, \cdots 1$ respectivement. Le paramètre degré N pour le noyau de Legendre et polynomial varie de 1 à 10. Les paramètres C, μ et N qui ont donné la meilleure exactitude SVM seront conservés, où

$$Exactitude = \frac{nombre\ de\ prédictions\ correctes\ données}{nombre\ total\ des\ données\ test} \times 100\%. \tag{4.13}$$

FIGURE 4.1 – Les données spirals

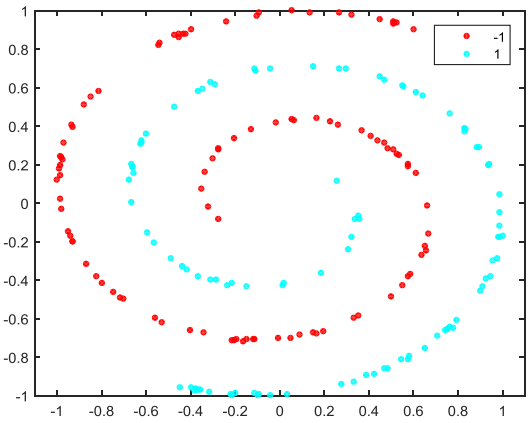


FIGURE 4.2 – 4X4 checkerboard (Damier)

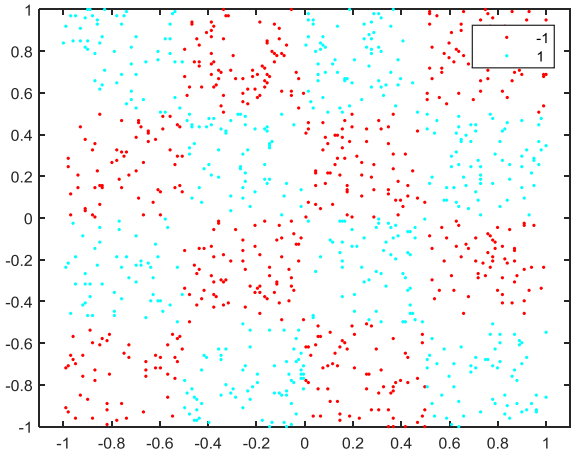


FIGURE 4.3 – Noyau RBF pour les données spirals

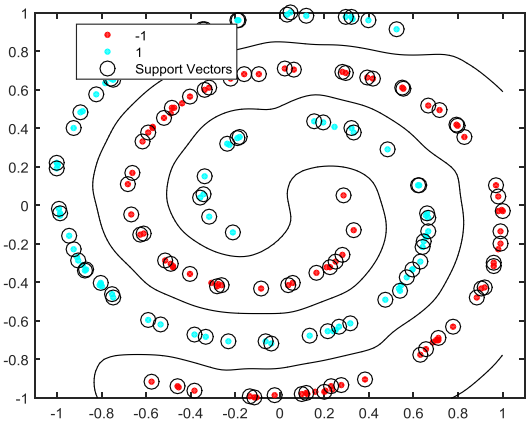
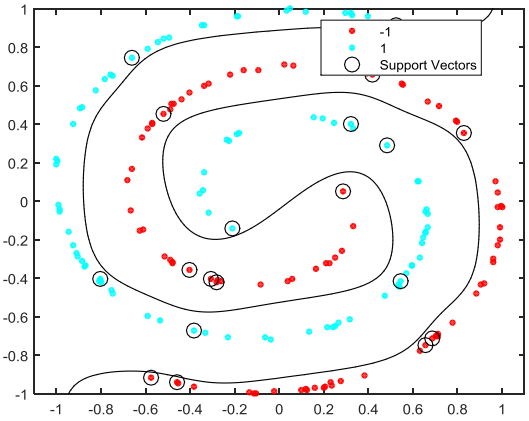


FIGURE 4.4 – Noyau de Legendre pour les données spiral



Un SVM final a été exécuté en utilisant les valeurs choisies de C, μ et N sur toutes les données d'exécutions. Le SVM résultant a été évalué sur les données test. La fonction noyau est utilisée pour calculer la matrice de Gram, dans "fitcsvm" de Matlab, RBF et Polynomial kernel sont disponibles en option, cependant l'utilisateur peut définir sa propre fonction noyau comme on l'a fait pour le noyau de Legendre. On n'a rapporté que les meilleurs résultats pour le noyau de Legendre, RBF et polynomial dans la figure (4.3, 4.4, 4.5, 4.6, 4.7 et 4.8). Dans le tableau (4.1), l'exactitude est l'exactitude du test, le temps est le temps de calcul écoulé pendant la phase d'apprentissage avec les meilleures valeurs de paramètres. On remarque un avantage en faveur du noyau de Legendre concernant le nombre de vecteurs Support.

4.2 Problèmes réels

Pour les problèmes réels de grande dimension sélectionnés dans le référentiel d'apprentissage automatique UCI, nous avons utilisé LIBSVM dans Matlab. Les mêmes étapes

FIGURE 4.5 – Noyau RBF pour les données de checker (Damier)

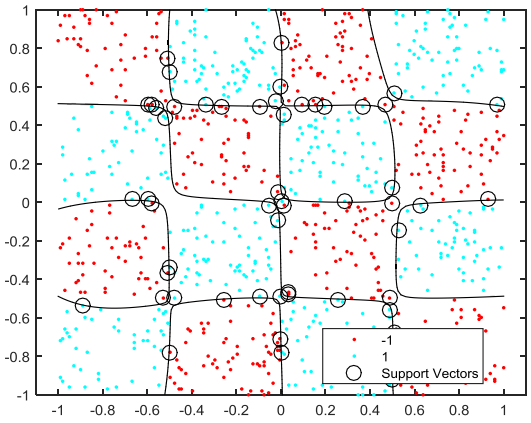


FIGURE 4.6 – Noyau de Legendre pour les données de checker (Damier)

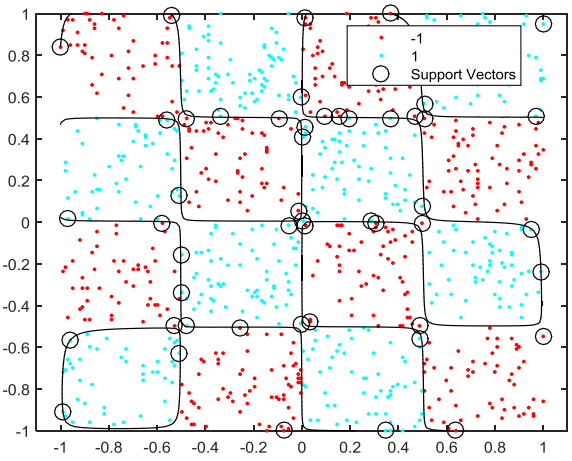


FIGURE 4.7 – Noyau polynomiale pour les données spirals

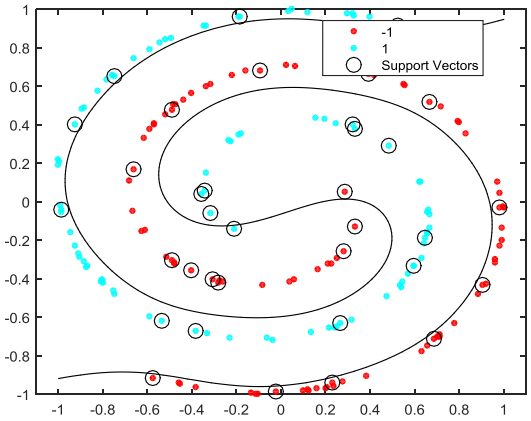
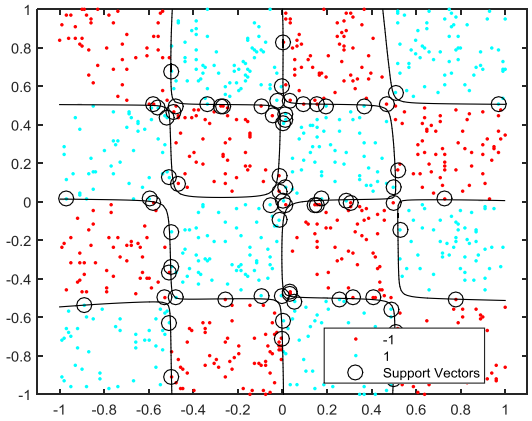


FIGURE 4.8 – Noyau polynomial pour les données du checker



de réglage des paramètres optimaux ont été utilisées et, comme les premiers tests, l'exactitude est l'exactitude des tests sur des échantillons non vus. Il est à noter qu'il faut plus de temps pour calculer la matrice de Gram des noyaux de Legendre et Polynomial que le noyau RBF. Cependant, LIBSVM a l'avantage d'introduire directement la matrice de Gram en entrée lorsque l'on utilise un noyau précalculé et le temps d'exécution du programme ne prend pas en compte le temps de calcul de cette matrice. Nous n'avons rapporté que les meilleurs résultats au tableau (4.2).

TABLEAU 4.1 – **Résultats numériques sur des exemples en 2 dimensions**

Exemples académiques (m,d)	Noyaux	Exactitude	Temps	Nombre des SVs	Les meilleurs paramètres	
Two spirals (200,2)	RBF	100%	0.0131	136	C=0.5	μ=0.1250
	Legendre	100%	0.4008	19	C=32	N=5
	polynômial	100%	0.0682	35	C=2	N=9
Checker (1000,2)	RBF	99%	0.9311	54	C=8192	μ=0.5
	Legendre	98%	8.8516	49	C=8192	N=5
	polynômial	95%	12.1141	70	C=8192	N=7

TABLEAU 4.2 – **Résultats numériques sur des exemples de grande dimension**

Problème UCI (m,d)	Noyau	L'exactitude	Temps	Nombre des SVs	Meilleurs paramètres	
Heart_scale (270×13)	RBF	92.5926%	0.0046	121	C=8	μ=2
	Polynômial	96.2963%	0.0027	103	C=0.0313	N=9
	Legendre	92.5926%	0.0039	217	C=0.0313	N=3
Diabetes (768×8)	RBF	70.1299%	0.0281	490	C=8192	μ=32
	Polynômial	76.6234%	0.0114	332	C=8	N=3
	Legendre	74.0260%	0.0279	396	C=0.1250	N=9
SPECTF (80×44)	RBF	75%	0.0263	66	C=2048	μ=2
	Polynômial	87.5%	0.0015	51	C=0.0313	N=1
	Legendre	87.5%	0.0022	69	C=0.0313	N=2
Breast_Cancer Wisconsin (683×10)	RBF	97.1014%	0.0158	334	C=2	μ=8
	Polynômial	98.5507%	0.0078	226	C=0.0313	N=7
	Legendre	98.5507%	0.0108	269	C=128	N=6
Bupa (345×6)	RBF	74.2857%	0.1930	216	C=32	μ=2
	Polynômial	74.2857%	0.0031	192	C=8	N=3
	Legendre	80%	0.0040	240	C=0.0313	N=9

Conclusion

Dans ce travail, on s'est intéressé au noyau de Legendre pour la classification non linéaire et on l'a revisité du point de vue mathématique. Essentiellement, on a montré qu'il n'est pas nécessaire d'avoir un RKHS avec une dimension infinie pour séparer les données dans l'espace caractéristique. Même les expériences numériques montrent que ce noyau peut rivaliser avec les noyaux universels RBF et Polynomial. Le seul inconvénient, c'est le temps nécessaire pour calculer la matrice de Gram associée au noyau de Legendre, en cas de grande dimension. Cependant, quand on voit le large spectre de polynômes orthogonaux pour lesquels le théorème (4.6) sera toujours applicable, on sera optimistes pour les travaux futurs. A tittre d'example, les polynômes orthogonaux de Jacobi, qui sont solutions de l'équation différentielles suivante

$$(1-x^2)y'' + (\beta - \alpha - (\alpha+\beta+2)x)y' + n(n+\alpha+\beta+1)y = 0, \quad y = \mathrm{J}_k^{\alpha,\beta}(x)$$

avec

$$\mathrm{J}_k^{\alpha,\beta}(x) = \frac{(-1)^k}{2^k k!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^k}{dx^k}[(1-x)^{k+\alpha}(x+1)^{k+\beta}]$$

Les polynômes de Jacobi recouvrent plusieurs cas particuliers : les polynômes de Legendre ($\alpha = \beta = 0$) et les polynômes de Tschebychef de premier, deuxième, troisième et quatrième espèces avec ($\alpha = \beta = \frac{-1}{2}$), ($\alpha = \beta = \frac{1}{2}$), ($\alpha = -\beta = \frac{-1}{2}$) et ($\alpha = -\beta = \frac{1}{2}$) respectivement. Le noyau de Jacobi a été déja introduit dans ([4]) où α et β étaient considéré comme des paramétre de réglage. Le théorème (4.6) nous informe, que l'orthogonalité joue un rôle fondamental. De ce fait, les familles d'ondelettes représentent aussi une bonne perspective.

En fin, les divers méthodes d'optimisation que nous avons étudié et qui ont été appliquées au problèmes quadratique des SVMs, nous ont montré que la technique de décomposition quand elle est associée aux techniques classiques de l'optimisation numérique, donnera naissance à des algorithme hybrides trés performants, cette voie mérite aussi beaucoup d'attention.

Bibliographie

[1] N. Aronszajn,"Theory of Reproducing Kernels", Transactions of the American Mathematical Society, Vol. 68, No. 3 , pp. 337-404 (1950) . 1

[2] K. Bache, and M. Lichman, "UCI Machine Learning Repository" University of California, Irvine, School of Information and Computer Sciences (2013).

[3] M.S. Bazaraa, H.D. Sherali et C.M. Shetty "Nonlinear Programming Theory and Algorithms", AJohn Wiley & Sons, INC.,Publication, (2006). 4, 5, 7

[4] N.Benmaghnia, "Noyaux à base de Polynomes orthogonaux pour les Machines à Vecteurs de Support (SVM). Mémoire de Master, option "Modélisaton, Controle et Optimisation". Mai 2017, Université Abdelhamid Ibn Badis, Mostaganem. 46

[5] C. Burges, "A tutorial on support vector machines for pattern recognition" , In Data-Mining and Knowledge Discovery . Kluwer Academic Publishers, Boston, (Volume 2) (1998). 1

[6] C-C. Chang, and C-J. Lin "LIBSVM : a library for support vector machines", ACM Transactions on Intelligent Systems and Technology,Vol. 2, No. 3, pp 1 :27(2011). 42

[7] N. Cristianini, and J.Shawe-Taylor, "An Introductin to Support Vector Macines and Other Kernel Based Learning Methods", Cambridge University Press. New York (2000). 29, 30, 32, 34

[8] I. Daubechies,"Ten Lectures on Wavelet", SIAM, CBMS-NSF regional conferences edition (1992).

[9] L.Debnath, and P. Mikusinki, "Introduction to Hilbert Spaces with applications' 3rd ed. San Diego", CA : Elsevier (2005).

[10] N. Djelloul, "Les Machines à Vecteurs de Support (SVMs). Mémoire de Master, option "Modélisaton, Controle et Optimisation". Juin 2012, Université Abdelhamid Ibn Badis, Mostaganem. 1

[11] N.Djelloul, and A.Amir, "Analysis of Legendre Kernel in Support Vector Machine" ,to appear in ' International Journal of Computing Science and Mathematics' (2018). 40, 41

[12] K.P Edwin Chang and H. Stanislaw Zak, "An Introduction to Optimisation' 2 nd ed, A Wiley-lnterscience Publication", (2001). iii, 4, 5, 10

[13] M.C. Ferris, O.L. Mangasarian et S.J. Wright "Linear Programming with MATLAB", MPS-SIAM Series on Optimization, (2007).

[14] S. Fine and K. Scheinberg," Efficient SVM training using low-rank kernel representations". The Journal of Machine Learning Research, 2 :243–264, (2002). 26

[15] G. Fung and O.L. Mangasarian " Breast Tumor Susceptibility to Chemotherapy via Support Vector Machines", (2006).

[16] L. Guillopé, Analyse fonctionnelle. "Approximations Hilbertiennes et développements en série". Ecole polytechnique de l'université de Nantes (2008) 37, 38

[17] L. Juntao, J. Yingmin and L. Wenlin "Adaptive Huberized Support Vector Machine and its Application to Microarray Classification", (2011).

[18] S.S. Keerthi and D. DeCoste. "A modified finite Newton method for fast solution of large scale linear SVMs". Journal of Machine Learning Research, (2006).

[19] Y.J. Lee and O.L. Mangasarian " Reduced Support Vector Machines", (2001). 26

[20] S. Lee and S.J. Wright. "Decomposition Algorithms for Training Large-Scale Semiparametric Support Vector Machines". In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases : Part II, page 14. Springer, (2009).

[21] M. Lichman "UCI Machine Learning Repository"[http ://archive.ics.uci.edu/ml]. Irvine, CA : University of California, School of Information and Computer Science (2013). 26

[22] O.L. Mangasarian " Arbitary-Norm Separating Plane", (1999). 42

[23] O.L. Mangasarian, "Nonlinear Programming", SIAM, Philadelphia, PA (1994).

[24] O.L. Mangasarian and D.R. Musicant. "Active support vector machine classification". Advances in Neural Information Processing Systems, 13 :577 – 583, (2000). 5

[25] E. Marchiori and M. Sebag "Bayesian Learning with Local Support Vector Machines for Cancer Classification with Gene Expression Data", (2005). 26

[26] MATLAB (2014) "Statistics and Machine Learning Toolbox Release (2014b)", *The MathWorks, Inc., Natick, Massachusetts, United States.*

[27] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations", Philos. Trans. Roy. Soc. London Ser. A vol 209 (1909). 42

[28] C.A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels", J. Mach. Learn. Res 7 :2651-2667 (2006).

[29] V.H. Moghaddam, and J. Hamidzadeh, "New Hermite orthogonal polynomial kernel and combined kernels in Support Vector Machine Classifier", Pattern Recognition 60 (921-935) (2016). 1

[30] J. Nocedal, and S. Wright, "Numerical Optimization", Springer-Verlag, New York (1999). 2

[31] S. Ozer, C.H.Chen, and H.A Cirpan, "A set of new Chebyshev kernel functions for support vector machine pattern classification", Pattern RecognitionVol 44, pp. 1435-1447(2011). 1, 8, 11, 13, 14, 26

[32] Z. PAN, H. CHEN, and X. YOU, "Support Vector Machine With Orthogonal Legendre Kernel", Proceedings of the 2012 International Conference onWavelet Analysis and Pattern Recognition, Xian 15-17 (2012). 2

[33] J. Platt. Sequential minimal optimization : "A fast algorithm for training support vector machines". Advances in Kernel Methods-Support Vector Learning, (1999). 2, 39

[34] A. Quarteroni, R. Sacco, F. Saleri, " Méthodes Numériques, Algorithmes, Analyse et Applications". Springer, Verlag Italia, Milano 2007. 26

[35] A. Rakotomamonjy, and S. Canu, "Frames, Reproducing Kernels, Regularization and Learning", Journal of Machine Learning Research 6, 1485-1515 (2005). 9

[36] R.A. Ryan, "Introduction to Tensor Products of Banach Spaces", Springer Monographs in Mathematics, Springer, London (2002). 38, 39

[37] K. Scheinberg, "An efficient implementation of an active set method for SVMs". Journal of Machine Learning Research, 7 :2237–2257, (2006). 2, 30, 31, 40, 41

[38] M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul (Eds.) "From Data and Information Analysis to Knowledge Engineering, Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V.", University of Magdeburg, Springer- Verlag (2005). 26

[39] I.Steinwart, "On the Influence of the Kernel on the Consistency of Support Vector Machines", Journal of Machine Learning Research 2 (2001). 1

[40] I. Steinwart and A. Christmann " Support Vector Machine", information Science and Statistic, Springer (2008). 31, 35, 36, 42

[41] S. Suvrit, S. Nowozin, and S.J. Wright, "Optimization for Machine Learning", Massachusetts Institute of Technology, London (2012). 29, 30, 31, 32, 33, 35

[42] J. Thongkam, G. Xu, Y. Zhang, and F. Huang "Support Vector Machine for Outlier Detection in Breast Cancer Survivability Prediction", (2008). 1

[43] R.J. Vanderbei "Linear Programming, Foundations and Extensions", Springer, (2008).

[44] V. Vapnik, "Estimation of Dependences Based on Empirical Data [in Russian]", Nauka, Moscow, 1979. (English translation : Springer Verlag, New York, 1982) (1979) . 6

[45] V. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, NewYork (1995). 1, 18, 26

[46] M. Vogt and V. Kecman. "Active-set methods for support vector machines". In Support Vector Machines : Theory and Applications. Springer Berlin Heidelberg, 2005. 18

[47] L. Wang, "Support vector machines : theory and applications", New York : Springer-Verlag (2005).

[48] T. Wen, A. Edelman, and D. Gorsich. "A fast projected conjugate gradient algorithm for training support vector machines". Joint Summer Research Conference on Fast Algorithms, pages 1–19, (2003). 1

[49] N. Ye, R. Sun, Y.Liu, and L. Cao, "Support vector machine with orthogonalnChebyshev kernel", in : Proceedings of the 18th International Conference on Pattern Recognition pp. 752-755 (2006). 26

[50] F. Zhou, Z. Fang, and J. Xu,"Constructing Support Vector Machine Kernels from Orthogonal Polynomials for Face and Speaker Verification", Fourth International Conference on Image and Graphics, IEEE (2007). 2

2